THE PHYSICAL SOCIETY

REPORTS ON PROGRESS IN PHYSICS

VOLUME XXIII (1960)

EXECUTIVE EDITOR
A. C. STICKLAND

Published by
THE PHYSICAL SOCIETY
1 Lowther Gardens, Prince Consort Road
London S.W.7
1960

Printed in Great Britain by John Wright and Sons Ltd.
The Stonebridge Press, Bath Road, Bristol 4

CONTENTS

	PAGE
PREFACE	iv
Optical Properties of Thin Films. By O. S. HEAVENS	1
Group Theory in Solid State Physics. By D. F. JOHNSTON	66
Photoelectric Photometry. By H. J. J. BRADDICK	154
Experimental Analysis of the Electronic Structure of Metals. By A. B. PIPPARD	176
New Developments in Interference Spectroscopy. By P. JACQUINOT	267
Planetary Nebulae. By M. J. SEATON	313
Band Structure Calculations in Solids. By L. PINCHERLE	355
Orbital Electron Capture by the Nucleus. By R. BOUCHEZ and P. DEPOMMIER	395
Precision Measurement in Gamma-ray Spectroscopy. By G. A. BARTHOLOMEW, J. W. KNOWLES and G. E. LEE-WHITING	453
^{3}He Induced Reactions. By D. A. BROMLEY and E. ALMQVIST	544

PREFACE

The *Reports on Progress in Physics* is under the general supervision of the Papers Committee of the Physical Society. The Members of the Papers Committee during the 1959–60 session were :

Chairman : Professor R. W. Ditchburn.

Honorary Papers Secretary : Dr. H. H. Hopkins.

Ordinary Members : The President (Mr. J. A. Ratcliffe), Dr. M. Blackman, Dr. R. L. F. Boyd, Professor W. E. Burcham, Dr. E. H. S. Burhop, Mr. G. J. Chamberlain, Dr. A. H. Cooke, Professor C. Domb, Professor B. H. Flowers, Professor F. C. Frank, Professor N. Kemmer, Dr. R. Latham, Dr. W. M. Lomer, Dr. K. Mendelssohn, Mr. K. H. Spring, Dr. R. W. B. Stephens, Professor D. H. Wilkinson, Dr. J. M. Ziman.

Reports published in the Volume are also being issued as individual articles, at prices depending on their length. Information may be obtained from the offices of the Society.

A. C. STICKLAND,
Secretary–Editor.

OPTICAL PROPERTIES OF THIN FILMS

By O. S. HEAVENS
Royal Holloway College, Englefield Green, Surrey

CONTENTS

PAGE

§ 1. Introduction ... 2

§ 2. Reflection of light at a single surface ... 3
2.1. Calculation of effect of surface layer ... 3
2.2. Experimental results ... 5

§ 3. Optical behaviour of a single film ... 6
3.1. General ... 6
3.2. Reflection and transmission at a single film ... 6
3.3. Normal incidence: non-absorbing media ... 8
3.4. Anisotropic and absorbing media ... 8

§ 4. Multilayer theory ... 10
4.1. Methods of calculation ... 10
4.2. Matrix method of solution ... 12
4.3. Graphical solutions of multilayer problems ... 13
4.4. General features governing the design of multilayer systems ... 15

§ 5. Measurement of film thickness and optical constants ... 16
5.1. General ... 16
5.2. Optical methods for measurements of the thickness and optical constants of films ... 16
5.2.1. Interferometric, fringe-displacement, methods ... 17
5.2.2. Photometric methods ... 19
5.2.3. Polarization methods ... 21
5.2.4. Methods employing stepped reflectors ... 23
5.2.5. Separate measurement of n and k ... 25
5.2.6. Brewster angle method for refractive index ... 27
5.2.7. Spectrophotometric methods of thickness and index determinations ... 28
5.3. Summary of methods available for optical constants and thickness determinations ... 28

§ 6. Results of measurements of optical constants of films ... 28
6.1. Metal films ... 30
6.1.1. Gold films ... 30
6.1.2. Silver films ... 35
6.1.3. Other metals ... 37
6.2. Dielectric films ... 41
6.2.1. Films formed by thermal evaporation ... 42
6.2.2. Chemically and electrolytically deposited films ... 47

§ 7. Methods of controlling film thickness ... 49
7.1. Semi-reflecting metal films ... 49
7.2. Dielectric layers ... 49
7.3. Piezoelectric crystal method ... 49

§ 8. Some applications of thin films in optics ... 50
8.1. Anti-reflecting systems ... 50
8.2. High-reflecting systems ... 51
8.2.1. Protected metal mirrors ... 51
8.2.2. Reflection filters ... 53
8.2.3. All-dielectric reflection systems ... 53

PAGE
8.3. Interference filters 55
8.3.1. Fabry–Perot type filter 55
8.3.2. Double and treble half-wave systems 56
8.3.3. The frustrated total reflection filter 58
8.4. Polarization devices 58
8.5. Heat-absorbing systems 59
Appendix I. Refractive indices of transparent materials in thin film form .. 60
Appendix II. Optical constants of opaque metal films 61
References 62

Abstract. The optical behaviour of a surface carrying a transition layer is discussed, as a problem in scattering. The results are given of the application of electromagnetic theory to the case of a single parallel-sided film. Methods are given for extending this treatment to the case of multiple parallel-sided layers. Both analytical and graphical solutions are considered. Features governing the design of certain multilayer systems are given. A survey is made of the many methods now available for the determination of the thickness and optical constants of materials in the form of thin films. The range, accuracy and conditions of application of the various methods are summarized. A selection of the results obtained on a range of metallic and dielectric films is made and the basis of interpretation of these results is given. It is shown that the apparently anomalous variations of the optical constants of metal films with film thickness can be attributed to the granular nature of such films. Although less spectacular variations are observed in the properties of dielectric films, it is seen that for some materials the optical behaviour suggests the presence of marked inhomogeneity and anisotropy. Films formed chemically and electrolytically are briefly mentioned.

Recent methods of controlling the thickness of films during deposition are given, together with a selection of the more important applications of thin films in optics. These include anti-reflecting and high reflecting systems, narrow-band, wide-band and cut-on multilayer filters, the use of thin films in polarizing systems and the application of thin films in heat-absorbing systems.

§ 1. Introduction

The optical behaviour of thin films may be considered from two viewpoints. From studies of the structures of the films formed by most available methods it is found that films generally consist of aggregates of crystallites of variable size and orientation and with differing extents of voids. Description of the optical behaviour of such a sytsem is essentially a scattering problem and this approach is often used in studying the effects of, for example, a transition layer on the surface of a solid. The mathematical problems involved are difficult generally but become tractable for layers whose thickness is of the order of molecular spacings.

It proves convenient to treat problems involving thicker layers—e.g. tens of millimicrons upwards—by describing the layer by a refractive index and, for absorbing materials, by an extinction coefficient. The optical properties of the layer can then be calculated using the classical electromagnetic theory. No difficulty arises in extending this approach to systems of multiple films: the reflectance and transmittance can be calculated explicitly in terms of the optical constants and film thicknesses. The mathematics is straightforward although cumbersome: problems are readily handled by digital computers or graphical methods may be found useful. The experimentally observed properties of films of some materials are found to agree closely with the predictions of electromagnetic theory, thus validating this approach to the film problem. For layers of thickness

of the order of molecular dimensions, however, scattering theory shows that such films cannot be described in terms of a refractive index and extinction coefficient. In cases where discrepancies are found for thick films these may be generally attributed to a lack of homogeneity of the film arising from, for instance, a change in film structure with film thickness. It is sometimes possible to deduce from the optical examination information on the film structure.

A wealth of methods has by now been developed for the measurement of the thickness and optical constants of films. Provided that the films are smooth and lie on a smooth substrate, interferometric and photometric methods enable a precision of better than 1 mμ to be attained. For transparent films the refractive index is directly measurable to within $\pm 0{\cdot}002$ under favourable conditions. The optical constants of absorbing films are generally rather more difficult to determine accurately : an accuracy to about 2% may be reached with existing methods. In the case of measurements on evaporated metal films, the sensitivity of the optical constants to conditions of preparation is so great that no higher accuracy in the optical measurement is at present called for.

With the increasing wide application of films in optical devices, methods of monitoring film thickness have been developed to an impressive standard. For the construction of multilayer filters of sophisticated design the precision of thickness control now attainable enables filters with (literally) scores of layers to be made. As more and more materials suitable for use in multilayer systems are discovered we move steadily towards the time when filters will be tailor-made for (almost) any desired purpose. No attempt is made in § 8 to give a comprehensive list of applications of thin films in optics : the examples discussed may be considered typical of the uses to which thin films may be put.

§ 2. Reflection of Light at a Single Surface

2.1. *Calculation of Effect of Surface Layer*

With the aid of electromagnetic theory the calculation of the reflectance and state of polarization of light reflected from the plane surface separating two homogeneous media is straightforward. The media are assumed to be continua and are described by refractive indices and, in the case of absorbing media, extinction coefficients. If visible light is considered, the assumption of a continuous medium is reasonably justified since the wavelengths entailed are some thousand times greater than the separations between atoms. For the case of reflection at the plane surface of an isotropic, homogeneous, non-absorbing medium of refractive index n, we obtain for the amplitude reflection coefficients r_p (electric vector $\parallel$ plane of incidence) and r_s (electric vector $\perp$ plane of incidence) :

$$r_p = -\frac{\tan(\phi_0-\phi_1)}{\tan(\phi_0+\phi_1)} \qquad \ldots\ldots(1)$$

$$r_s = +\frac{\sin(\phi_0-\phi_1)}{\sin(\phi_0+\phi_1)} \qquad \ldots\ldots(2)$$

where ϕ_1, ϕ_0 are the angles of incidence and refraction respectively. The sign convention is that used by Ditchburn (1952). The difference between the phase

changes on reflection $\delta = \delta_p - \delta_s$ undergoes an abrupt change at the polarizing angle ϕ_p, this being defined as the angle of incidence at which $\phi_0 + \phi_1 = \frac{1}{2}\pi$. Figure 1 shows the calculated dependence of δ on angle of incidence together with a typical experimental curve. The lack of sharpness of the experimentally observed change of δ with ϕ_1 is ascribed to a transition layer at the surface. This may arise from, for instance, the action of polishing, the presence of contamination or effects of surface tension on the atomic arrangement near the surface. Since the sensitivity attainable by standard polarimetric techniques is high, very small departures from the theoretical prediction for the homogeneous solid can be detected. Such methods therefore provide a powerful means of studying surface films, provided the observations can be correctly interpreted.

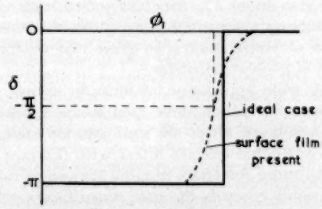


Figure 1. Dependence of δ on angle of incidence ϕ.

The effect of a continuous variation of the optical properties of a surface layer on the state of reflected light was treated by Drude (1891) and McLaurin (1905). Strachan (1933) treated the case of a two-dimensional distribution of dipoles on a surface, obtaining expressions for the reflection coefficients and phase changes in terms of three scattering coefficients, defined as the components of the strengths of the Hertzian dipoles produced at the surface of the medium by the electric field of the incident wave. The results are not too intractable for an isotropic film. Considerable simplification is effected by the treatment of Sivukhin (1956) in which it is shown that the effect of a transition layer of any thickness can be reduced to that of two infinitesimally thin layers with dipole moment surface densities. For a transition layer of thickness small compared with the wavelength of the light used, the problem reduces to that of considering a single surface dipole moment distribution.

The treatment is restricted to the calculation of the intensities of reflected waves at distances large compared with the intermolecular distances in the medium. That this approximation is entirely satisfactory is shown by Ewald (1916): the radiation field from a crystal lattice is found to be practically equal to that for a continuous medium even at distances from the surface of the order of the lattice constant of the crystal.

For a plane monochromatic wave incident on the surface of the medium, the polarization in the medium may be written:

$$\mathbf{P} = \mathbf{P}_0 \exp i(\omega t - \kappa r). \qquad \ldots\ldots(3)$$

The effect of a thin transition layer (thickness $\ll \lambda$) may be expressed in terms of a supplementary dipole moment per unit area of surface, given by

$$\tau = \tau_0 \exp i(\omega t - \kappa_x x) \qquad \ldots\ldots(4)$$

where Oz is normal to the surface and $y = 0$ is the plane of incidence. If it is assumed that the supplementary polarization at the surface is small compared with the polarization in the absence of the transition layer, then the results for the reflection coefficients may be expressed in terms of parameters γ_m $(m = x, y, z)$ given by

$$\gamma_m = \frac{\tau_{0m}}{P_{0m}}. \qquad \ldots\ldots(5)$$

The expressions for the reflection coefficients r_p and r_s may then be written

$$r_p = -\frac{\tan(\phi_0 - \phi_1)}{\tan(\phi_0 + \phi_1)}\left[1 + 2if\cos\phi_0 \frac{\gamma_x \cos^2\phi_1 - \gamma_z \sin^2\phi_0}{\cos^2\phi_1 - \sin^2\phi_0}\right] \qquad \ldots\ldots(6)$$

and

$$r_s = \frac{\sin(\phi_0 - \phi_1)}{\sin(\phi_0 + \phi_1)}[1 + 2if\gamma_y \cos\phi_0]. \qquad \ldots\ldots(7)$$

f is the magnitude of the wave vector of the incident wave. For an isotropic surface distribution, $\gamma_x = \gamma_y$.

The results given above are obtained from a first-order approximation based on the assumptions stated. It is shown (Sivukhin 1956) that in taking approximation to the second order, it is necessary to consider quadrupole radiation from the transition layer. In principle if the molecular structure of the surface is known, the parameters γ_m (and also the further parameters of the second-order approximation) are calculable.

An interesting result of the first-order theory, and one which does not appear to have been put to experimental test, is that the ellipticity of the light reflected from a surface covered by a transition layer should be inversely proportional to the wavelength of the radiation used. A further conclusion is that a transition can be characterized phenomenologically by a refractive index only if its thickness is large compared with molecular dimensions. Under this condition, the first-order Sivukhin equations reduce to those of Drude whilst the second-order equations yield the McLaurin results.

2.2. *Experimental Results*

Experimental investigation of the transition layer on the surfaces of solids suffers from the difficulty that, in order that accurate measurements can be made, an optically smooth, flat surface is required. There appears to be no way of producing such a surface without using a method which is liable to produce surface disorder. Measurements of the differential phase change on (total) reflection at a glass prism face (Ditchburn and Orchard 1954) show large differences from the value calculated from the (bulk) refractive index, even for a range of different surface treatments. The effect of repolishing the surface was to yield results corresponding to a transition layer about 10 mμ thick. Measurements by the same authors on light totally reflected from a flowing water surface were found to agree closely with the predictions of simple theory, suggesting that the results obtained

on solids may well arise from unavoidable surface disorder. The effect of surface tension forces may still, however, cause departures from simple theory in the case of a solid since the estimated surface tensions in solids are 1–2 orders of magnitude higher than those observed in liquids.

§ 3. Optical Behaviour of a Single Film

3.1. *General*

Much interest has centred in recent years on films, or systems of films, of thickness comparable with the wavelengths of visible light. The practice of coating lenses with a quarter-wave anti-reflecting (blooming) layer is now common. If multiple reflections are neglected, it is a trivial calculation to show that, for complete extinction at a given wavelength, the refractive index of the anti-reflecting layer n_a must equal the square root of the substrate index. In fact, this result is still correct if the effect of multiple reflections is included. Practical difficulties prevent the easy realization of the index required for coating a typical crown glass. The low index needed can be achieved only by arranging that the film be full of holes : the mechanical properties of such films are unsatisfactory. Magnesium fluoride is widely used : although its index is higher than the optimum, it may be deposited as a very hard, stable, durable film. Results of recent studies of the properties of MgF_2 are given in § 6.2.

3.2. *Reflection and Transmission at a Single Film*

Expressions for the reflection and transmission coefficients of a surface covered with a single homogeneous isotropic film are conveniently expressed in terms of the Fresnel reflection and transmission coefficients for the two interfaces. These quantities may conveniently be defined here in general terms since they are needed for the treatment of multiple layers in § 4. The Fresnel coefficients (ratio of reflected or transmitted to incident amplitude) depend on the plane of polarization considered. For light incident from a medium of refractive index n_m on the boundary with a medium of index n_{m-1} we have (figure 2) for the p-component†

$$r_{mp} = \frac{n_m \cos\phi_{m-1} - n_{m-1}\cos\phi_m}{n_m \cos\phi_{m-1} + n_{m-1}\cos\phi_m} \quad \text{(reflection)} \qquad \ldots\ldots(8)$$

$$t_{mp} = \frac{2n_m \cos\phi_m}{n_m \cos\phi_{m-1} + n_{m-1}\cos\phi_m} \quad \text{(transmission)}. \qquad \ldots\ldots(9)$$

For the s-component

$$r_{ms} = \frac{n_m \cos\phi_m - n_{m-1}\cos\phi_{m-1}}{n_m \cos\phi_m + n_{m-1}\cos\phi_{m-1}} \qquad \ldots\ldots(10)$$

$$t_{ms} = \frac{2n_m \cos\phi_m}{n_m \cos\phi_m + n_{m-1}\cos\phi_{m-1}}. \qquad \ldots\ldots(11)$$

The ϕ's are related by Snell's law :

$$n_m \sin\phi_m = n_{m-1}\sin\phi_{m-1}. \qquad \ldots\ldots(12)$$

† The definitions given here for Fresnel coefficients differ from those quoted in some publications (e.g. Abelès 1948 a, Heavens 1955). The reason is simply that, for some treatments of the multi-layer problem, it proves more convenient to label the substrate n_0 and number the films upwards from the substrate. In order that the notation for various treatments given below shall be unified, this system will be adopted throughout.

In order to simplify the equations, Fresnel coefficients will generally be written simply as r_m, t_m, with the understanding that when (8) and (9) are used, the result gives the p-component and, with (10) and (11), the s-component.

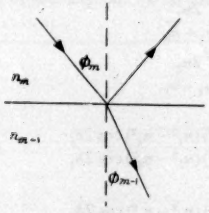


Figure 2.

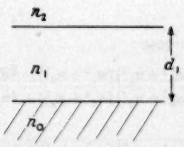


Figure 3. Single film.

For a single film of index n_1 and thickness d_1 on a substrate of index n_0 (figure 3), the amplitude reflection and transmission coefficients are given by

$$R = \frac{r_2 + r_1 \exp(-2i\delta_1)}{1 + r_1 r_2 \exp(-2i\delta_1)} \qquad \ldots\ldots(13)$$

$$T = \frac{t_1 t_2 \exp(-i\delta_1)}{1 + r_1 r_2 \exp(-2i\delta_1)} \qquad \ldots\ldots(14)$$

where
$$\delta_1 = 2\pi\nu n_1 d_1 \cos\phi_1 \qquad \ldots\ldots(15)$$

and
$$\nu = 1/\lambda.$$

These results follow simply and directly from summation of multiple beams in the film. For non-normal incidence equations (13) and (14) each yield two equations, corresponding to each plane of polarization.

The energies crossing unit area per second normal to the reflected and transmitted beams are given by $n_0 RR^*$ and $n_2 TT^*$. Hence the reflectance of the system is given by

$$\mathbf{R} = RR^* = \frac{r_1^2 + 2r_1 r_2 \cos 2\delta_1 + r_2^2}{1 + 2r_1 r_2 \cos 2\delta_1 + r_1^2 r_2^2}. \qquad \ldots\ldots(16)$$

If the transmittance $\mathbf{T}$ is defined as the ratio of illumination on screens perpendicular to the transmitted and incident beams, then

$$\mathbf{T} = \frac{n_2}{n_0} TT^* = \frac{n_2}{n_0} \frac{t_1^2 t_2^2}{1 + 2r_1 r_2 \cos 2\delta_1 + r_1^2 r_2^2}, \qquad \ldots\ldots(17)$$

We may note that the transmission characteristic may equally well be defined by the ratio of the total flux transmitted by unit area of the film to that incident (Weinstein 1954) in which case the transmittance $\mathbf{T}_f$ would be

$$\mathbf{T}_f = \frac{n_2 \cos\phi_2}{n_0 \cos\phi_0} \frac{t_1^2 t_2^2}{1 + 2r_1 r_2 \cos 2\delta_1 + r_1^2 r_2^2}, \qquad \ldots\ldots(18)$$

For the case of normal incidence, $\mathbf{T} = \mathbf{T}_f$.

3.3. *Normal Incidence: Non-absorbing Media*

Equations (16) and (17) may be written directly in terms of the refractive indices of the media. The Fresnel coefficients are equal for both planes of polarization, and are given by

$$r_1 = \frac{n_1 - n_0}{n_1 + n_0}, \quad t_1 = \frac{2n_1}{n_1 + n_0}, \qquad \ldots\ldots(19)$$

$$r_2 = \frac{n_2 - n_1}{n_2 + n_1}, \quad t_2 = \frac{2n_2}{n_2 + n_1}. \qquad \ldots\ldots(20)$$

For this case

$$\mathbf{R} = \frac{(n_0^2+n_1^2)(n_1^2+n_2^2) - 4n_0 n_1^2 n_2 + (n_0^2-n_1^2)(n_1^2-n_2^2)\cos 2\delta_1}{(n_0^2+n_1^2)(n_1^2+n_2^2) + 4n_0 n_1^2 n_2 + (n_0^2-n_1^2)(n_1^2-n_2^2)\cos 2\delta_1} \qquad \ldots\ldots(21)$$

$$\mathbf{T} = \frac{8n_0 n_1^2 n_2}{(n_0^2+n_1^2)(n_1^2+n_2^2) + 4n_0 n_1^2 n_2 + (n_0^2-n_1^2)(n_1^2-n_2^2)\cos 2\delta_1}. \qquad \ldots\ldots(22)$$

These equations are seen to be symmetrical with respect to n_0 and n_2, showing that the values of reflectance and transmittance given above apply for light incident from either side of the film.

Equation (22) gives the intensity of the beam in the substrate medium n_0. If the light emerges from the back of the substrate medium, then allowance must be made for any reflection effect at the surface of emergence. If the substrate is a parallel-sided slab, multiple reflections must be considered. If the substrate thickness is large, so that multiply reflected beams are incoherent, intensities are added. If the substrate thickness is small, so that the beams are coherent then the system becomes a two-film system and the treatment of § 3 must be used.

Equations (13) and (14) may still be used for the case of absorbing media: the values of n used in evaluating the Fresnel coefficients and in the δ_1 term are replaced by a complex $\mathbf{n} \equiv n - ik$ where k is the extinction coefficient. The values of ϕ_1 obtained from Snell's law are complex, a manifestation of the fact that, in an absorbing medium, the planes of constant amplitude may be inclined to the planes of constant phase of the wave.

The determination of optical constants of an absorbing film from reflection and transmission data, using a digital computer, is described by Harris and Loeb (1955).

3.4. *Anisotropic and Absorbing Media*

The case of an anisotropic, absorbing film sandwiched between transparent isotropic media has been dealt with by Schopper (1952 a). The more general case in which any or all of the three media are anisotropic and absorbing has been treated by Bousquet (1957) in the following way.

It is first necessary to evaluate the Fresnel coefficients for the case of reflection and transmission at the interface between two such media. For the coordinate system shown in figure 4 we may write for the phase terms of the electric vectors of the incident, reflected and refracted waves:

$$\exp i\omega(t - \alpha_1 x - \gamma_1 z), \quad \exp i\omega(t - \alpha_1 x + \gamma_1 z)$$

and

$$\exp i\omega(t - \alpha_2 x - \gamma_2 z).$$

For an isotropic, non-absorbing system, $\alpha_i = (n_i \sin\phi_i)/c$ and $\gamma_i = (n_i \cos\phi_i)/c$ where $i = 1, 2$ and ϕ is the angle of incidence. For the general case, the value of n_i appropriate to the direction of propagation must be inserted. Bousquet's treatment is restricted to the case in which the normal to the surface is a privileged direction of the crystal. It is then possible to express the propagation terms

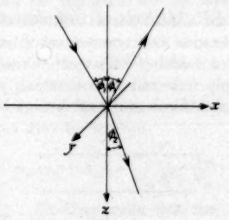


Figure 4.

α and γ quite simply in terms of the principal dielectric constants $\epsilon_x, \epsilon_y, \epsilon_z$. For absorbing media these are complex. The Fresnel coefficients may be written

(i) Electric vector $\perp$ plane of incidence

$$r_{1p} = \frac{\gamma_1 - \gamma_0}{\gamma_1 + \gamma_0}, \quad t_{1p} = \frac{2\gamma_1}{\gamma_1 + \gamma_0} \qquad \ldots\ldots(23)$$

where

$$\gamma_0^2 = \frac{\epsilon_{y0}}{c^2} - \alpha_1^2. \qquad \ldots\ldots(23a)$$

(ii) Electric vector $\parallel$ plane of incidence

$$\left.\begin{aligned} r_{1s} &= \frac{\epsilon_{x1}\gamma_0 - \epsilon_{x0}\gamma_1}{\epsilon_{x1}\gamma_0 + \epsilon_{x0}\gamma_1} \\ t_{1s} &= \frac{2\gamma_1/\epsilon_{x1}}{(\gamma_1/\epsilon_{x1} + \gamma_0/\epsilon_{x0})} \end{aligned}\right\} \qquad \ldots\ldots(24)$$

where

$$\gamma_0^2 = \frac{\epsilon_{x0}}{c^2} - \alpha_1^2 \frac{\epsilon_{x0}}{\epsilon_{z0}}. \qquad \ldots\ldots(24a)$$

The values of r_{1p}, t_{1p} and r_{1s}, t_{1s} may thus be obtained in terms of the direction of the incident beam and of the dielectric constants of the media. Expressions for $r_{2p}, t_{2p}, r_{2s}, t_{2s}$ for the case of incidence from medium 2 to medium 1 may be written down by inspection from equations (23) and (24). The amplitude reflection coefficients for the case of a film of thickness d_1 are then obtained from

$$R = \frac{r_2 + r_1 \exp(-2i\omega\gamma_1 d_1)}{1 + r_1 r_2 \exp(-2i\omega\gamma_1 d_1)} \qquad \ldots\ldots(25)$$

$$T = \frac{t_1 t_2 \exp[-i\omega d_1(\gamma_1 - \gamma_2)]}{1 + r_1 r_2 \exp(-2i\omega\gamma_1 d_1)} \qquad \ldots\ldots(26)$$

where values of r, t appropriate either to the p- or s-components of polarization are inserted. With the help of equations (23*a*) and (24*a*), together with their

counterparts for the 2–3 interface, R and T may be calculated in terms of (α_2, γ_2) and the parameters of the media. For absorbing media, the explicit expressions are cumbersome: the computation is best done stepwise.

§ 4. Multilayer Theory

4.1. *Methods of Calculation*

Calculation of the reflectance and transmittance of a system of parallel-sided films has been effected in a number of ways. Crook (1948) has extended the method of summing multiply reflected beams, which yields the result for a single film easily, to the case of two films and has outlined the form of procedure for

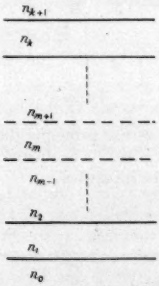


Figure 5. Notation for multilayer system.

expressing the result for n layers in terms of that for $n-1$ layers. The process is however cumbersome and not convenient for computation. Rouard (1937) and Vašíček (1950) make use of the fact that, since a single film may be characterized by an effective reflection coefficient and phase change, then such a film may be replaced, for the purpose of computing reflectances, by a single surface with these properties. By a stepwise procedure, one can work through the whole system of layers until the final boundary is reached. For a system of k layers on a substrate of index n_0 (figure 5), the amplitude reflection coefficient for the first film may be written

$$\rho_2 \exp(i\Delta_2) = \frac{r_2 + r_1 \exp(-2i\delta_1)}{1 + r_1 r_2 \exp(-2i\delta_1)} \qquad \ldots\ldots(27)$$

where r_2, r_1 are the Fresnel coefficients for the n_2/n_1 and n_1/n_0 interfaces respectively. The reflection coefficient for the first two films may then be written

$$\rho_3 \exp(i\Delta_3) = \frac{r_3 + \rho_2 \exp(i\Delta_2) \exp(-2i\delta_1)}{1 + r_3 \rho_2 \exp(i\Delta_2) \exp(-2i\delta_1)}. \qquad \ldots\ldots(28)$$

Thus $\rho_4, \Delta_4, \rho_5, \Delta_5, \ldots \rho_{k+1}, \Delta_{k+1}$ are determined successively. The amplitude reflection coefficient for the whole system of films is ρ_{k+1}, so the reflectance is simply ρ_{k+1}^2. The extension to the case of absorbing layers simply entails replacing the real indices n by complex values $n - ik$ in the expressions for r and δ. It should

be noted however that the simple concept of multiple reflections in an absorbing layer forming part of the multilayer system may give rise to difficulties (Berning 1956).

A disadvantage of this method of calculation is that if the effect of altering the thickness or refractive index of any layer in the system is being studied, then the procedure given above must be repeated for all layers subsequent to the one changed. This drawback is avoided in the matrix method discussed below.

Some simplification results for the case of normal incidence and recurrence relations have been given by Godfrey (1957) which are convenient for obtaining some features of systems of practical importance.

For the system of figure 5, and for normal incidence, the amplitude reflection coefficient of the first m films may be written

$$\rho_m \exp(i\Delta_m) = \frac{n_{m+1} - N_m}{n_{m+1} + N_m} \qquad \ldots\ldots(29)$$

which serves to define N_m. Godfrey calls this the 'film index'—perhaps not a satisfactory name. The recurrence relation between N_{m-1} and N_m is readily seen to be

$$\frac{N_m - n_m}{N_m + n_m} = \frac{N_{m-1} - n_m}{N_{m-1} + n_m} \exp(-4\pi i \nu n_m d_m). \qquad \ldots\ldots(30)$$

For absorbing layers n_m is replaced by $\mathbf{n}_m \equiv n_m - ik_m$. If none of the layers is absorbing the reflectance and transmittance of the system of k layers are given by

$$\mathbf{R}_k = \left| \frac{n_{k+1} - N_k}{n_{k+1} + N_k} \right|^2 \qquad \ldots\ldots(31)$$

$$\mathbf{T}_k = \frac{4 n_0 n_{k+1}}{(n_1 + n_0)^2} \prod_1^k \left| \frac{N_m + n_m}{N_m + n_{m+1}} \right|^2. \qquad \ldots\ldots(32)$$

These expressions are particularly convenient for calculating the maximum reflectance of quarter-wave stacks. For this case the exponential term in equation (30) becomes unity, thus simplifying the recurrence relation for N_m, N_{m-1}. For an even number $2s$ of layers of the form LHL ... LH, the reflectance is given by

$$\mathbf{R}_{2s} = \left(\frac{n_0 f - n}{n_0 f + n} \right)^2 \qquad \ldots\ldots(33)$$

where $f = (n_H/n_L)^{2s}$ and n is the refractive index of the medium of incidence. For the system of $2s+1$ layers, HLH ... LH, the reflectance is

$$\mathbf{R}_{2s+1} = \left(\frac{f n_H^2 - n n_0}{f n_H^2 + n n_0} \right)^2. \qquad \ldots\ldots(34)$$

For high-reflecting stacks, in which the reflectance approaches unity, these may be written in the convenient approximate forms

$$\mathbf{R}_{2s} \simeq 1 - \frac{4n}{n_0 f}; \quad \mathbf{R}_{2s+1} \simeq 1 - \frac{4 n n_0}{n_H^2 f}. \qquad \ldots\ldots(35)$$

The special case in which the thicknesses of the layers forming a periodic multilayer system are small compared with the wavelength used has been dealt with by Vasilkovsky (1957). Such a system behaves as a single film with an

equivalent refractive index equal to the root-mean-square of the indices of the layers of the system. The accuracy of the calculations using the simplified formula improves as the ratio thickness/wavelength decreases. For semiconducting materials the simple formula is valid for thickness/wavelength not greater than 10^{-2}.

4.2. *Matrix Method of Solution*

The equations relating the amplitudes of the electric vectors in successive media can be conveniently written in matrix notation (Weinstein 1947, Herpin 1947, Abelès 1948 b, Muchmore 1948, Canals Frau 1952, Sukhanovsky 1956, Kard 1957 a). This form is generally the most convenient for computation either by hand or by digital computer.

If we denote by E^+_m the amplitude of the electric vector of the wave travelling in the direction of incidence in the mth layer and by E^-_m that of the negative-going wave, solution of the equations of propagation readily yields the following relations between $E^\pm_m$ and $E^\pm_{m-1}$:

$$\left.\begin{aligned} E^+_m &= t_m^{-1}[E^+_{m-1}\exp(i\delta_{m-1}) + r_m E^-_{m-1}\exp(-i\delta_{m-1})] \\ E^-_m &= t_m^{-1}[r_m E^+_{m-1}\exp(i\delta_{m-1}) + E^-_{m-1}\exp(-i\delta_{m-1})]. \end{aligned}\right\} \quad \ldots\ldots(36)$$

Where the r_m and t_m are given,† for whichever plane of polarization is considered, by equations (8) to (11).

Equation (36) may be written

$$\begin{aligned} \begin{pmatrix} E^+_m \\ E^-_m \end{pmatrix} &= \frac{1}{t_m}\begin{pmatrix} \exp(i\delta_{m-1}) & r_m\exp(-i\delta_{m-1}) \\ r_m\exp(i\delta_{m-1}) & \exp(-i\delta_{m-1}) \end{pmatrix}\begin{pmatrix} E^+_{m-1} \\ E^-_{m-1} \end{pmatrix} \\ &= \frac{1}{t_m} M_{m-1}\begin{pmatrix} E^+_{m-1} \\ E^-_{m-1} \end{pmatrix}. \end{aligned} \quad \ldots\ldots(37)$$

The solution for a stack of k layers on a substrate is thus given simply by

$$\begin{pmatrix} E^+_{k+1} \\ E^-_{k+1} \end{pmatrix} = \frac{M_k M_{k-1}\ldots M_2 M_1}{t_k t_{k-1}\ldots t_2 t_1}\begin{pmatrix} E^+_0 \\ E^-_0 \end{pmatrix} \quad \ldots\ldots(38)$$

from which the reflectance $\mathbf{R} = \dfrac{|E^-_{k+1}|^2}{|E^+_{k+1}|^2}$

is readily obtained. The transmittance $\mathbf{T}$ is given by

$$\left.\begin{aligned} &\text{p-component:} && \mathbf{T}_p = \frac{n_0}{n_{k+1}}\frac{|E^+_0|^2}{|E^+_{k+1}|^2} \\ &\text{s-component:} && \mathbf{T}_s = \frac{n_0\cos^2\phi_{k+1}}{n_{k+1}\cos^2\phi_0}\frac{|E^+_0|^2}{|E^+_{k+1}|^2}. \end{aligned}\right\} \quad \ldots\ldots(39)$$

The value of the transmittance defined in terms of light fluxes [equation (18)] is given, for both p- and s-components, by

$$\mathbf{T}_f = \frac{n_0\cos\phi_{k+1}}{n_{k+1}\cos\phi_0}\frac{|E^+_0|^2}{|E^+_{k+1}|^2}. \quad \ldots\ldots(40)$$

† We note the existence of needless confusion of notation here. Although the notation for Fresnel coefficients consistent with that given in equations (8)–(11) has been generally used (Cotton 1950, Malé 1950, Mayer 1950) the notation used by Weinstein (1954) differs, although not sufficiently to avoid the danger of confusion.

In evaluating the latter expression, however, the appropriate p- or s-values of Fresnel coefficients must be used.

An alternative form of recurrence relation may be obtained in terms of the components of **E** and **H**, giving

$$\begin{pmatrix} E_{m+1} \\ H_{m+1} \end{pmatrix} = \begin{pmatrix} \cos\delta_m & (i/u_m)\sin\delta_m \\ iu_m \sin\delta_m & \cos\delta_m \end{pmatrix} \begin{pmatrix} E_m \\ H_m \end{pmatrix} \qquad \ldots\ldots(41)$$

where $E_m = E^+{}_m + E^-{}_m, \quad H_m = u_m(E^+{}_m - E^-{}_m).$

For the p-component $u_m = n_m \cos\phi_m$ and for the s-component $u_m = n_m/\cos\phi_m$. One important advantage of this form of equation (41) as compared with equation (37) is that the square matrix of equation (41) contains only quantities characteristic of one layer of the system. The Fresnel coefficients in equation (37) depend on the indices of two media. It will be noted that the determinant of the square matrix is unity. Since the determinant of the product of any number of such matrices is also unity, the correctness of computation is easily checked at any stage. For non-absorbing films, the leading diagonal elements are real and the product determination straightforward. If absorbing films are present, then equation (41) offers no particular advantages over equation (36) for purposes of computation.

The development of experimental techniques in recent years enables systems of dozens of layers to be made without difficulty. Hand computation for such problems is entirely uneconomic. This type of computation is easily handled by an automatic digital computer, which can be programmed to deal with the problem more cheaply, more reliably and more accurately than by the use of human labour.

4.3. *Graphical Solutions of Multilayer Problems*

If the scale of work contemplated does not justify the programming of a computer, or if computer services are not available, there are several graphical methods available for dealing with multilayer problems (Malé 1950, Cotton 1950, Blaisse 1950, Krylova 1958). One of the best known is taken from the field of transmission lines, in which the Smith Chart (Kraus 1949) has long been in use. Using the variable $y_m = H_m/u_m E_m$ (known as the admittance), the recurrence relation for y is

$$y_{m+1} = \frac{u_m}{u_{m+1}} \frac{i\sin\delta_m + y_m \cos\delta_m}{\cos\delta_m + iy_m \sin\delta_m} \qquad \ldots\ldots(42)$$

which may be written in terms of the amplitude reflection coefficient

$$R_m = E^-{}_m/E^+{}_m$$

as

$$y_{m+1} = \frac{u_m}{u_{m+1}} \left(\frac{1 - R_m \exp(-2i\delta_m)}{1 + R_m \exp(-2i\delta_m)} \right). \qquad \ldots\ldots(43)$$

The coordinate mesh of the Smith Chart is that of the function y of the conformal transformation $y = (1-R)/(1+R)$ so the chart may be used to obtain R_{m+1} from R_m and hence, by successive use, the amplitude reflection coefficient for the whole system.

A slightly more convenient form of graphical calculator is that of Kard (1956). For the interface between the $(m+1)$th and mth layers, the amplitude reflection coefficient may be written

$$R_{m+1} = \frac{r_{m+1} + R_m \exp(-2i\delta_m)}{1 + r_{m+1} R_m \exp(-2i\delta_m)} \qquad \ldots\ldots(44)$$

or

$$R_{m+1} = \tanh \tfrac{1}{2}(A_m + B_m) \qquad \ldots\ldots(45)$$

where we write

$$A_m = 2 \operatorname{arc\,tanh} r_{m+1} \qquad \ldots\ldots(46)$$

$$B_m = 2 \operatorname{arc\,tanh} [R_m \exp(-2i\delta_m)]. \qquad \ldots\ldots(47)$$

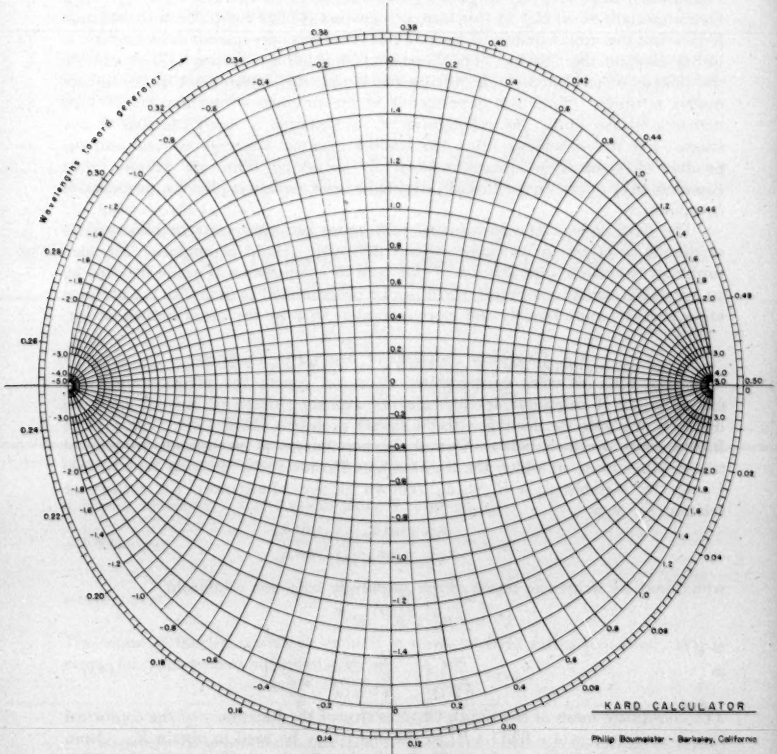


Figure 6. Kard's calculator.

The Kard calculator† (figure 6) is a mapping of the conformal transformation $z = \tanh \frac{1}{4}w$ which thus enables A_m, B_m and R_{m+1} to be determined from given values r_{m+1} and $R_m \exp(-2i\delta_m)$. If all the layers are non-absorbing then the values of A_m [equation (46)] are all real. For normal incidence these simply take the form

$$A_m = 2 \operatorname{arc\,tanh}\left(\frac{n_{m+1}-n_m}{n_{m+1}+n_m}\right) = \ln\left(\frac{n_m}{n_{m+1}}\right). \qquad \ldots\ldots(48)$$

4.4. *General Features governing the Design of Multilayer Systems*

The use of the digital computer for solving multilayer problems has been mentioned above (§ 4.2). Such methods have reached an impressive degree of sophistication in which the computer, on being presented with a trial multilayer system and with data for the desired reflectance or transmittance characteristic, refines the parameters of the multilayer in order to achieve the desired result. It is, however, necessary that the trial system with which the computer is provided shall possess characteristics fairly close to the ideal. For this reason, methods which enable the general features of a multilayer system to be predicted may well serve a useful purpose.

Epstein (1952) has examined the properties of some symmetrical systems of films, making use of the general results of the treatment of Herpin (1947) that a symmetrical system may be reduced to a single film with a certain equivalent index and thickness. These quantities may be readily obtained from the parameters of the system: Epstein gives results for three-layer systems. The transmission characteristics of the system are found to be divided between pass bands, in which the equivalent index is real, and high-reflectance (' stop ') bands, for which the equivalent index is imaginary. In the pass-band the equivalent thickness is approximately equal to the real thickness. The behaviour of the system in these bands is therefore easily visualized in terms of the (known) properties of a single layer.

In an alternative approach Smith (1958) shows that a multilayer system may be reduced to two interfaces. This representation is not confined to symmetrical systems. (The reflection properties of the representative interfaces may be such that they would not be shown by a single equivalent film: this is of no consequence since the behaviour of the system is deduced from the reflectance characteristics of the interfaces and not from equivalent film constants.) The transmission characteristic of a system with two interfaces is easily represented in terms of the Airy sum, and the dependence of such characteristics on the reflectances at the two interfaces is easily visualized. Successive application of this approach enables the properties of various more complex systems to be qualitatively deduced. Thus the properties of systems of mixed quarter- and half-wave layers are readily predicted. In application of the method to these systems, the conditions necessary for narrow-band, wide-band and low-pass (in frequency) filters are readily obtained. The properties of some of these systems are discussed in § 8.

† The copy of the Kard calculator is reproduced by kind permission of Dr. P. W. Baumeister of Rochester, N.Y.

§5. Measurement of Film Thickness and Optical Constants

5.1. *General*

The method used for the determination of the thickness of a film depends on a large number of factors. The film may be transparent or absorbing, as may also be the substrate. The edge of the film may be accessible, enabling interference step methods to be used; in this case, however, care is needed if part of the film lies below the surrounding surface as in the case of a metallic oxide. The choice of method depends on the order of the thickness involved: we shall deal with methods suitable for films up to thickness of the order of microns.

It must be mentioned that in certain cases the film thickness—or, rather, the quantity yielded by the methods of measurement given below—may not be a very well-defined physical quantity. In the case of aggregated films, in which the diameters of aggregates may exceed the mean thickness $\bar{t}$ (defined, for example, as $A^{-1}\Sigma t\,\delta A$ where A is the area), the mass per unit area is perhaps the only useful parameter. The mass per unit area of films of the order of a centimetre square may be determined directly by microbalance (Clegg and Crook 1952, Bradley 1953), with which films of down to about 1 mμ may be handled. Microchemical techniques exist for the estimation of microgramme quantities of many elements (Sandell 1944, Johnson 1955) which in favourable cases enable the mass of 1 cm^2 × 10 mμ films to be determined to better than about 0·5% (Heavens 1959). These methods have the drawback that the film is destroyed in the measurement process. Direct thickness measurement using the 'Talysurf' instrument enables measurements of the order of 20 mμ to be made within $\pm$2 mμ; this method is suitable for hard, coherent films (soft films may be damaged by the stylus). Electrical resistance measurements are in principle sensitive enough for thickness measurements down to a few tenths of a millimicron, but the resistivity of such films is generally so critically dependent on the conditions of preparation, nature of substrate, age, etc., that such methods can be used only with great circumspection. The experiments of Alderson and Ashworth (1957) show, however, that reproducible results may be obtained with nickel–chromium films of thicknesses up to a few tens of millimicrons provided the conditions of preparation are adequately standardized. With the many other electrical measurements which depend on film thickness—electrolytic methods, work function measurements, capacity measurements—the dependence of the electrical properties on thickness needs first to be determined; this may be a major project. More direct methods are to be preferred.

5.2. *Optical Methods for Measurements of the Thickness and Optical Constants of Films*

Several methods are available for determining thickness and/or optical constants of films, the general requirement being that the films have a smooth surface and lie on a smooth substrate. The methods may be classified roughly as follows.

Interferometric methods. Certain of these methods utilizing fringe displacements, give the film thickness only, independently of the optical constants of film and substrate. Others entail the location of the wavelengths of interference maxima and minima; these require a knowledge of the optical properties of the substrate.

Photometric methods. Measurements of the reflectance at each side of a film on a transparent substrate can yield the thickness of the film and the optical constants simultaneously, albeit with considerable labour.

Polarized light methods. A study of the state of polarization of light reflected from a film-covered surface may yield information on the film thickness and optical constants. The quantities generally determined are (i) the difference Δ in the phase change on reflection for the two components of polarization and (ii) the ratio of the amplitudes of the two reflected components. In the case of films on transparent substrates, measurements of reflectance and transmittance may be combined with phase change measurements.

5.2.1. *Interferometric, fringe-displacement, methods.*

One of the simplest methods of measuring film thickness is that using multiple-beam interference fringes (Tolansky 1948). The edge of the film is coated with an opaque silver film and measurements are made on the shift in the reflection fringes (Fizeau fringes) observed with a monochromatic source. Interference takes place between the silvered surface and a heavily silvered reference flat. If the order of interference can be made very low, the fringes are extremely fine and an accuracy in a typical case of ± 1 mμ can be obtained.

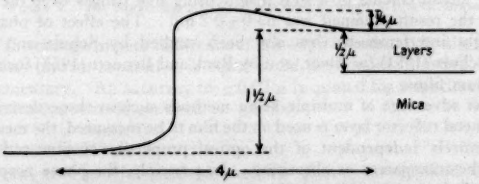


Figure 7. Distortion at edge of multilayer reflecting stack.

It has been established (Heavens 1951, Weaver and Benjamin 1956) that, provided an opaque reflecting layer is used and there is no chemical or other deterioration of the surfaces, the method yields the true thickness of the underlying film. The thickness of the reflecting layer is unimportant, so long as it is opaque: accurate contouring is reported for thicknesses of a silver coating up to 5 μ (Tolansky 1951). The use of high-reflecting multilayers for this purpose, in place of silver films, has been studied (Belk, Tolansky and Turnbull 1954): the lower absorption attainable in a dielectric multilayer is an advantage. Comparison of thickness measurements using silver and multilayer reflectors shows close agreement. There is a slight tendency to distortion (figure 7) of the contouring of the multilayer at the extreme edge of the step.

Fringes which are inherently sharper than Fizeau fringes and which are formed by illuminating a parallel-sided interferometer with white light and then dispersing the reflected (or transmitted) light may be used for thickness measurement (fringes of equal chromatic order). Although in principle a greater accuracy is attainable with these latter fringes than with Fizeau, it is necessary to apply corrections before an accurate result is obtained. Since white light is used, and measurements

are made on fringes of different wavelengths, correction for the phase change on reflection at the silver (or multilayer) surfaces must be applied. In view of the variation between different observers' results for measurement of phase changes on evaporated metal films, arising from the dependence of this quantity on the precise conditions of preparation of the film, this correction cannot easily be applied. Experiments by Scott, McLaughlan and Sennett (1950) suggest that phase change effects are negligible: the step heights deduced from fringes of orders 1 to 30 are found to be the same even if no phase change effect is assumed. This is, in fact, slightly strange since if the phase change correction is calculated using published values of optical constants, there should be a most marked influence over the range of wavelengths used. In a similar arrangement, systematic errors in thickness measurements are reported by Koehler (1953). When multilayers are used as reflectors the effect of phase changes on reflection is much larger than with a metal reflector and must be allowed for. The procedure to be followed is given by Koehler (1955). The phase change on reflection is not required explicitly but may be eliminated by suitable treatment of the data. In fact, Koehler additionally calculated the effect of phase changes from the known constants of the multilayers used, by means of equation (41), and obtained good agreement over the range of ± 100 mμ from the wavelength at which the peak multilayer reflectance occurs. In measurements of a step height using five fringes over the range 460 to 560 mμ, the result obtained was $63{\cdot}0 \pm 0{\cdot}2$ mμ. The effect of phase changes in white light interferometry has also been studied by Schulz and Scheibner (1950), by Schulz (1951) for silver, and by Rank and Bennett (1955) for multilayers and aluminium films.

The great advantage of multiple-beam methods such as those described above is that, if a metal reflector layer is used on the film to be measured, the measurements made are entirely independent of the optical properties of film and substrate. Either may be transparent or absorbing. In principle, the phase properties of a multilayer will depend on the optical constants of the substrate. For a multilayer with many layers on a transparent substrate, the dependence on substrate properties will be slight.

The use of Fabry–Perot and Michelson interferometers for film thickness measurement of transparent films have been described by Schulz (1950) and Fochs (1950). In Schulz's method, the film is deposited over part of an optical flat which is then heavily silvered and used as one mirror of a Fabry–Perot interferometer, illuminated with sodium light. There are then two successive positions of the mirror, separated by a distance M mm, in which the Fabry–Perot fringes appear as shown in figure 8. The film thickness may be determined from M; the wavelength difference between the sodium lines is used as an internal standard. The film thickness t in mμ is given by $t = 507M$. The setting of figure 8 can be made to within a tenth of a fringe, corresponding to $\pm 1{\cdot}5$ mμ in film thickness. The possible effect of a different substrate on the phase change on reflection of the silver film is not discussed although the effect for high-reflecting silver films is probably negligible.

Both film thickness and refractive index of a (non-absorbing) film are obtained by the use of the Michelson interferometer with a white light source. The film is placed so as to intercept part of one of the two beams. Part of one mirror is

covered by an opaque screen (figure 9) so that the fringes seen in the dispersing system from this region (A) arise from multiple reflections in the film. Normal Michelson fringes are seen from region C whilst those from B are displaced on

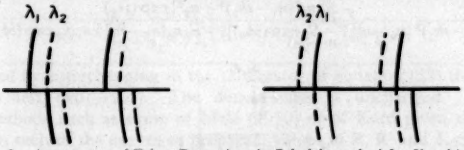


Figure 8. Appearance of Fabry–Perot rings in Schulz's method for film thickness.

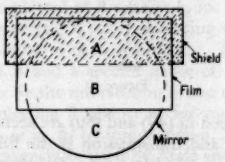


Figure 9. The Michelson interferometer used to measure film thickness.

account of the change in path difference of $2(n-1)d$ introduced by the film. When the Michelson mirror separation is equal to $2d$ the fringes from regions A and B are complementary. An accuracy to $\pm 0{\cdot}02\ \mu$ is quoted for a specimen 12 μ thick. The refractive index is determined to $\pm 0{\cdot}003$.

5.2.2. *Photometric methods.*

There are two general methods in this class: the method described by Schopper (1952 b) entails measurement of the amplitude and phase of the light reflected from each side of a film and of the light transmitted by the film, all the measurements being at (or as near as is possible to) normal incidence. Although three measurements should suffice to determine the three unknowns n_1, k_1, d_1, the nature of the equations relating reflectance and transmittance to these quantities is such that n_1, k_1 and d_1 cannot be written explicitly in terms of only three measurables.

Denoting by S_i, ϕ_i the amplitudes and phase changes on reflection in media n_0, n_2 (figure 10), S, ϕ the amplitude and phase change on transmission, and writing

$$\epsilon_2 = \phi_2 - \pi + \frac{4\pi n_2 d_1}{\lambda} \qquad \ldots\ldots(49)$$

and

$$\epsilon = \phi + \frac{2\pi n_2 d_1}{\lambda}, \qquad \ldots\ldots(50)$$

the index $n_1 - ik_1$ of the film is obtained in terms of the observables as

$$(n_1-ik_1)^2 = n_0 \frac{n_0^2 S_2 \exp(i\epsilon_2)[1+S_0\exp(i\phi_0)]^2 - n_2 S^2 \exp(2i\epsilon)\{n_2 - n_0[1+S_0\exp(i\phi_0)]\}}{n_0 S_2 \exp(i\epsilon_2)[1-S_0\exp(i\phi_0)]^2 + S^2\exp(2i\epsilon)\{n_0 - n_2[1-S_0\exp(i\phi_0)]\}} \qquad \ldots\ldots(51)$$

from which n_1 and k_1 are determined. The film thickness is then obtained from

$$\exp\left(\frac{4\pi i n_2 d_1}{\lambda}\right) = \frac{-S_2 n_0[(n_1-ik_1)^2-n_2^2]\exp(i\epsilon_2)}{(n_1-ik_1)^2\{n_0-n_2[1-S_0\exp(i\phi_0)]\}+n_2 n_0\{n_2-n_0[1+S_0\exp(i\phi_0)]\}}. \quad \ldots\ldots(52)$$

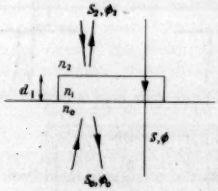


Figure 10.

The quantities ϵ, ϵ_2 introduced in (49) and (50) are seen to be the differences in the phase changes on reflection and transmission at the filmed and uncoated parts of the substrate. These are more readily measurable than the ϕ values.

This method possesses the advantage that (i) the thickness and optical constants are determined simultaneously and (ii) monochromatic light is used so that no errors from neglected dispersion effects arise. The disadvantage is that rather elaborate experimental arrangements are needed in order to make the (absolute) reflectance, transmittance and phase change measurements with sufficient accuracy. The accuracy of measurements made on films of up to 200 mμ thickness was estimated as to ± 3 mμ on thickness and to $\pm 0{\cdot}03$ on the optical constants.

Measurements of phase changes are avoided in the method introduced by Malé (1952), developed from an earlier method by Murmann (1933, 1936), in which the thickness and optical constants are determined from the measured reflectances and transmittance. Since it is impossible to express n_1, k_1 and d_1 explicitly in terms of the measured quantities the following indirect method is used. Sets of curves are first prepared giving **R**, **R′** (reflectances at either side of the film) and **T** as functions of $\delta \equiv 4\pi n_1 d_1/\lambda$ for a (two-dimensional) range of values of n_1 and k_1. These may be calculated from equations (13) and (14) by inserting $n_1 - ik_1$ in the expressions for the Fresnel coefficients. In closed form, for a film on a non-absorbing substrate of index n_0, the expressions for **R**, **R′** and **T** are

$$\mathbf{R} = \frac{(g_2^2+h_2^2)\exp(2\alpha_1)+(g_1^2+h_1^2)\exp(-2\alpha_1)+2(g_1g_2+h_1h_2)\cos 2\gamma_1+2(g_2h_1-g_1h_2)\sin 2\gamma_1}{\exp(2\alpha_1)+(g_1^2+h_1^2)(g_2^2+h_2^2)\exp(-2\alpha_1)+2(g_1g_2-h_1h_2)\cos 2\gamma_1+2(g_2h_1+g_1h_2)\sin 2\gamma_1} \quad \ldots\ldots(53)$$

$$\mathbf{T} = \frac{n_0}{n_2}\,\frac{[(1+g_1)^2+h_1^2][(1+g_2)^2+h_2^2]}{\exp(2\alpha_1)+(g_1^2+h_1^2)(g_2^2+h_2^2)\exp(-2\alpha_1)+2(g_1g_2-h_1h_2)\cos 2\gamma_1+2(g_1h_2+g_2h_1)\sin 2\gamma_1} \quad \ldots\ldots(54)$$

where $\alpha_1 = 2\pi k_1 d_1/\lambda, \qquad \gamma_1 = 2\pi n_1 d_1/\lambda,$

$$g_1 = \frac{n_1^2 - n_0^2 + k_1^2}{(n_0+n_1)^2 + k_1^2}, \qquad h_1 = \frac{-2n_0 k_1}{(n_0+n_1)^2 + k_1^2},$$

$$g_2 = \frac{n_2^2 - n_1^2 - k_1^2}{(n_2+n_1)^2 + k_1^2}, \qquad h_2 = \frac{2n_2 k_1}{(n_2+n_1)^2 + k_1^2}.$$

$\mathbf{R}'$ is obtained by interchanging in the numerator of equation (53) the coefficients of $\exp(2\alpha_1)$ and $\exp(-2\alpha_1)$. The denominator is unchanged. Alternatively graphical methods such as those of Malé (1950) or of Kard given above may be used. From each of the curves so prepared, values of $\mathbf{R}$, $\mathbf{R}'$ and δ, corresponding to the experimentally observed value of $\mathbf{T}$, may be obtained for various values of n_1 and k_1. Curves are then plotted of $\mathbf{R}$ against k_1 for each value of n_1 from which are read off pairs of values of (n_1, k_1) corresponding to the observed value of $\mathbf{R}$. These points are plotted on a curve of k_1 against n_1. The whole procedure is then repeated for $\mathbf{R}'$ in place of $\mathbf{R}$ and a second curve of k_1 against n_1 plotted. The required n_1 and k_1 are given by the intersection of the two curves. With n_1 and k_1 known, d_1 is obtained by plotting δ against n_1 for various k_1 and deducing the δ corresponding to the values of n_1, k_1 by interpolation.

This method has been extensively used by Malé and his collaborators in studies of the properties of metal films formed by thermal evaporation. The results of these experiments are discussed in § 6.1.

5.2.3. *Polarization methods.*

For a considerable time use has been made of the fact that measurements of the state of polarization of light beams may be made with high accuracy. Light reflected from a surface is generally elliptically polarized: by compensating the phase difference Δ between the p- and s-reflected components, plane polarized light is produced. The angle ψ ($= \arctan R_p/R_s$) between the plane of the electric vector and the plane of incidence is accurately measurable. The values of Δ and ψ depend sensitively on the presence of a film on the surface. In contrast to the interferometric methods given in § 5.2.1 however, the values of the optical constants of the substrate must be known. These are determined by standard polarimetric procedures (see, for example, Bor and Chapman 1949, Winterbottom 1946) on an uncoated part of the substrate. Alternatively, where substrates may be prepared bearing stepped films, the optical constants of the underlying metal may be eliminated. This method is discussed in § 5.2.4 below.

Transparent films.

(i) *Drude's approximation.* As early as 1889, Drude gave expressions, valid for films of thickness $d \ll \lambda$, enabling the thickness and refractive index to be determined from the measured differences in Δ and ψ for coated and uncoated parts of the substrate. The validity of these equations has been examined by Rothen and Hanson (1949) for films of barium stearate on stainless steel, close agreement being obtained for thicknesses up to 25 mμ.

(ii) *Vašíček's method.* If the Drude method is extended to deal with films of any thickness, it is no longer possible to deduce the values of index and thickness

directly from the equations. Vašíček (1947) has dealt with the problem of a transparent film on a glass substrate, giving a procedure whereby the values of n and d can be obtained. With the notation of §3.2, the values of Δ and ψ are readily seen to be given by

$$\tan\Delta = (AB' + AB')/(AA' - BB') \qquad \ldots\ldots(55)$$

where

$$A = r_{1p}(1 + r_{2p}^2) + r_{2p}(1 + r_{1p}^2)\cos 2\delta_1$$

$$A' = r_{1s}(1 + r_{2s}^2) + r_{2s}(1 + r_{1s}^2)\cos 2\delta_1$$

$$B = -r_{2p}(1 - r_{1p}^2)\sin 2\delta_1$$

$$B' = r_{2s}(1 - r_{1s}^2)\sin 2\delta_1$$

and

$$\tan^2\psi = \frac{(r_{1p}^2 + r_{2p}^2 + 2r_{1p}r_{2p}\cos 2\delta_1)(1 + r_{1s}^2 r_{2s}^2 + 2r_{1s}r_{2s}\cos 2\delta_1)}{(1 + r_{1p}^2 r_{2p}^2 + 2r_{1p}r_{2p}\cos 2\delta_1)(r_{1s}^2 + r_{2s}^2 + 2r_{1s}r_{2s}\cos 2\delta_1)}. \qquad \ldots\ldots(55a)$$

In Vašíček's experimental arrangement plane polarized light is incident with its plane of polarization inclined at 45° to the plane of incidence. The ellipticity of the reflected light is measured using a Senarmont compensator and Nakamura double plate (Vašíček 1940).

Equations (55) and (55a) cannot be solved directly for n_1 and d_1 but are dealt with by successive approximation. Vašíček gives a table of values of ψ and Δ for films of thickness up to 1–1·5 μ and for refractive indices 1·2 to 2·75 for a substrate of index 1·5163 and for $\lambda = 589$ mμ. Vašíček's claim that the method is sufficiently sensitive to detect films 1 mμ thick has been challenged by Mattuck (1956). It would appear that Vašíček's claim implies a precision of 0·5′ in the determination of Δ, ψ. In Mattuck's experimental arrangement, a precision of $\pm 2{\cdot}5'$, corresponding to a limiting film thickness of 5 mμ, was recorded.

Figure 11. Thickness measurement using the Françon eyepiece (van Heel).

(iii) *Use of the Savart plate.* An accuracy to $\pm\lambda/200$ can be attained in the measurement of the refractive index and thickness of a transparent film on a transparent substrate by the use of the eyepiece devised by Françon (1952, 1953, 1954). Plane polarized light from the polarizer P_1 (figure 11) emerges from the film and substrate with the wave front deformed as shown at A. The Savart plate in the eyepiece produces two mutually perpendicularly polarized component wave fronts which are displaced laterally with respect to each other (B). Viewed through the analyser P_2 the field of view will then in general show three different intensities, if monochromatic light is used, or three different colours, if white light is used. The use of monochromatic light is to be preferred unless the effects of dispersion can be adequately allowed for. Various experimental arrangements are discussed by van Heel and Walther (1958) for determining the phase difference introduced by

the film and the transmittance of the film from which, with the help of equation (14), the film thickness and refractive index are determined. Since normal incidence on the film is used and since no silvering is involved, the method is particularly suitable for, for instance, the continuous monitoring of the thickness of a film during deposition.

(iv) *Zernike–Dyson method.* This arrangement, in which high sensitivity is achieved by the use of a half-shade system, is one of the most sensitive to date for the measurement of the metrical thickness of a film. The edge of the film must be accessible and the surface is aluminized. Light from a slit is projected, via a Wollaston prism, on to the film surface, producing two images at the surface which are perpendicularly polarized. On re-passing the Wollaston prism the images are combined and the final image is examined through a polarizer and half-shade system. The dividing edge of the latter lies across the slit image. Settings for the equality of brightness are made (i) with the two slit images astride the film edge and (ii) with both images on one side of the film edge. Special arrangements are necessary to ensure that the setting adjustments do not alter the conditions of polarization. A repeatability to $\pm 0{\cdot}1$ mμ is obtained (Dyson 1958) in a system in which no serious precautions were taken against mechanical or thermal drifts.

Absorbing films.

The method due to Vašíček given above may be used to deal with absorbing films if the thickness is determined independently. Tables are given (Vašíček 1951) enabling the optical constants to be determined from the observations for the following ranges :

Thickness	to 100 mμ by 10 mμ
n	1·00 to 2·75 by 0·25,
k	0 to 1·0 by 0·2

Equations enabling the optical constants and film thickness to be deduced from measurements of the ellipticity of light transmitted by, and reflected from, each side of a film have been given by Försterling (1937). No approximations are involved. When this method was applied, complex values were obtained for the film thickness. This is likely to be a manifestation of anisotropy in the film and suggests that this method may be unduly sensitive to a departure from isotropy or homogeneity of the film. Since anisotropy in films generally takes the form of an optic axis normal to the substrate, methods involving only normal incidence measurements are to be preferred.

For dealing with absorbing films of such thickness and absorption that multiple reflections in the film may be neglected, Leberknight and Lustman (1939) use approximations which are found to be valid, for several metal oxides, for thicknesses up to several tens of millimicrons.

5.2.4. *Methods employing stepped reflectors.*

These methods make use of the fact that monomolecular layers of certain fatty acids and their salts may be easily picked up from a water surface. Multiple layers of such films are found to be stable and to behave accurately reproducibly in respect of their physical properties. Barium stearate has been used extensively.

Its properties have been very thoroughly investigated by Blodgett and Langmuir (1937). Such films may be built up to any number of layers. Their optical properties were fully investigated, using polarimetric and Brewster angle techniques, from which the thickness per film was deduced as 2·44 mμ.† The films are found to be uniaxial, with optic axis normal to the plane of the film, and with refractive indices $n_1 = 1{\cdot}491$ (ordinary) and $n_3 = 1{\cdot}551$ (extraordinary) for $\lambda = 589$ mμ.

Use is made of such films in an ingenious arrangement by Rothen (1945) for the determination of the optical thickness of films of thicknesses down to a few tenths of a millimicron. Different numbers of BaSt layers are picked up on two halves of a polished chromium slide. The slide is illuminated with plane polarized light which is reflected through a mica $\lambda/4$ plate so arranged that elliptically polarized light (of very small ellipticity) from one half of the field is left-handed and that of the other half right-handed. An analyser is adjusted for equality of brightness in the two fields. When a further film is deposited over the whole slide the brightnesses of the two fields become unequal in the manner obtaining in the usual half-shade system. Measured rotation of the analyser restores the uniformity of brightness and the amount of rotation yields the film (optical) thickness. When a BaSt layer was used as the added film, the rotation of the analyser was 3·2°. Since a rotation of 0·02° is detectable and since the thickness of one BaSt layer is 2·4 mμ the noise level of the measurement is of the order of 0·03 mμ. It must be stressed that this method yields the *optical* thickness and that the refractive index must be determined independently if the metrical thickness is required.

It is suggested by Hartman (1954) that the above procedure is liable to be in error if the index of the added film differs from that of the substrate. If the objection is valid the method becomes severely limited in scope. The grounds for the objection are not, however, made clear and no substantiation is given. The method proposed by Hartman in the same paper is the subject of some criticism by Mattuck (1956), who claims that the approximations used are not valid for films thicker than about 1 mμ.

For the study of films of the BaSt type, which may be stacked on a substrate in known numbers, Hartman *et al.* (1954) employ an interferometric method in which white light, plane polarized with the electric vector perpendicular to the plane of incidence, is incident on the surface bearing a suitable number of films (21 for BaSt). The reflected light is passed through an analyser set at 90° to the plane of incidence and is passed to a dispersing system. The s-component of the amplitude reflectance of the filmed surface is given by

$$R_s = \frac{r_{1s} + r_{2s}\exp(-2i\delta_1)}{1 + r_{1s}r_{2s}\exp(-2i\delta_1)} \qquad \ldots\ldots(56)$$

and will be a minimum for some values $2\delta_1 = C$, say, and for other values $2\delta_m$ such that

$$2\delta_m = C + 2m\pi \qquad \ldots\ldots(57)$$

where m is an integer. (The result holds only if dispersion of the optical constants of film and substrate are neglected. These assumptions are reasonably justified for BaSt on chromium for visible wavelengths.)

† Subsequent measurements give the slightly lower value of 2·28 mμ.

If d_0 is the thickness of one BaSt double layer and N is the number of double layers on the slide, then for an angle of incidence ϕ :

$$m = \frac{2Nd_0(n_1^2 - \sin^2\phi)^{1/2}}{\lambda} - \frac{C}{2\pi} \qquad \ldots\ldots(58)$$

where n_1 is the index of the film. A plot of m, the order of the dark fringe observed in the spectrograph, against N/λ yields a straight line of slope $b_1 = 2d_0(n_1^2 - \sin^2\phi)^{1/2}$. From this relation,

$$\sin^2\phi = -\frac{b_1^2}{4d_0^2} + n_1^2. \qquad \ldots\ldots(59)$$

Measurements are made at several angles of incidence (ranging from 70° to 85°) and the value of d_0 obtained from a plot of $\sin^2\phi$ against b_1^2. The refractive index is obtained from the slope of the line. From these experiments, the values of d_0 and n_1 (which corresponds to the ordinary index) are $d_0 = 4{\cdot}57 \pm 0{\cdot}07$ mμ, $n = 1{\cdot}508 \pm 0{\cdot}013$.

This method has been further developed by Mattuck, Petti and Bateman (1956). A stepped interference reflector consisting of two BaSt steps (of 20 and 40 double layers) is used. Fringes are formed in the focal plane of a dispersing system as in the Hartman method. Measurements are made of the displacements of fringes from both halves of the field when an unknown layer is deposited over the whole of the slide carrying the stepped reflector. For refractive indices in the range 1·1 to 2·0 and for thicknesses from 2 mμ to 25 mμ a sensitivity to 1% in refractive index and 4% in thickness is claimed. It is suggested that errors arising from neglect of birefringence in the BaSt films and of dispersion may be negligible. The values obtained for the refractive indices of several evaporated films (Mattuck, Petti and Bateman 1956) were found to agree within the experimental error with those obtained by the Abelès method (see §5.2.6 below). In general, the accuracy of the Abelès method is higher ; this method does not, however, give the film thickness.

5.2.5. *Separate measurement of n and k.*

Whilst the simultaneous measurement of n and k on one film eliminates any variation between different samples, it is a fact that n and k can be more accurately determined by separate experiments. Thus k may be accurately determined by transmission measurements, or by measurements of the phase change on reflection at a surface. With the value of k known, n is accurately obtained from measurements of the reflectance at the surface. These methods have been developed by Schulz and Tangherlini (1954) in experimental studies relating to the Drude free electron theory of metals. Measurements were made on metal films formed by thermal evaporation and annealed. The criterion for homogeneity adopted is the relation between the reflectances of the p- and s-components predicted by electromagnetic theory. For reflection at the interface between homogeneous media at an angle of incidence of 45°, $\mathbf{R}_s^2 = \mathbf{R}_p$. This relation was found to be satisfied by suitably annealed films of silver, aluminium and copper.

Calculation of the transmission of films of various thickness yields curves such as those of figure 12. The departure from linearity at small thicknesses arises

from the effect of reflections at the interfaces. When the thickness is sufficiently large the effect of multiple reflections in the film becomes negligible : under this condition the transmittance is given by

$$\mathbf{T} = \frac{16n_0 n_2(n_1^2+k_1^2)^2}{[(n_1+n_2)^2+k_1^2][(n_0+n_1)^2+k_1^2]} \exp\left(-\frac{4\pi k_1 d_1}{\lambda}\right). \quad \ldots\ldots(60)$$

From this equation we see that k_1 can be obtained from the slope of the curve of ln $\mathbf{T}$ against d_1. (Although in principle the intercept of this curve could yield n_1, using the k_1 value from the slope, Schulz suggests that the k value in the intercept term is likely to differ from that in the exponential term, on account of surface defects.)

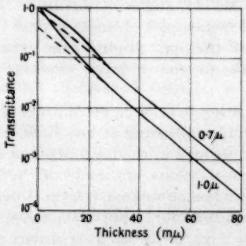


Figure 12. Calculated transmittance–thickness curves for silver films. The measured optical constants at $\lambda = 0{\cdot}7\ \mu$ and $\lambda = 1{\cdot}0\ \mu$ are used in the calculation.

An interesting alternative method for k is to make an interference filter and to deduce the value of the phase change α on reflection at the metal–dielectric interface from the effect of such phase change on the wavelengths of the transmission maxima, when white light is incident on the filter. k is then calculated from the equation

$$\tan\alpha = \frac{2k_1 n_0}{n_1^2+k_1^2-n_0^2} \quad \ldots\ldots(61)$$

where n_0 is the refractive index of the dielectric layer. The method requires (i) that the dispersion of the dielectric is known and (ii) that the metal film is thick enough for equation (61) (which holds for reflection at a bulk surface) to be valid. Furthermore, if the wavelengths of the transmission maxima are to be accurately measurable, the fringes must be sharp and hence the reflectivity of the metal films used must be high. This would seem to limit the method to metals with a large reflectivity and small ratio of absorption to transmission.

Measurement of the value of n is effected by measuring the reflectance at 45° of a thick (> opaque) film of the metal. In order to avoid the danger of atmospheric contamination, the reflection at the substrate side is measured (figure 13). A simple approximation method enables n to be obtained, using the value of k obtained, for example, in the transmission experiment reported above.

Since these methods were evolved with a view to studying the values of n and k, the film arrangements have been designed to suit the experiments rather than vice

versa. It is likely that the values of n and k obtained in these experiments are the most reliable to date.

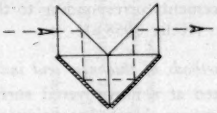


Figure 13. Arrangement for measuring the reflectance, at 45°, at the film–substrate interface.

5.2.6. *Brewster angle method for refractive index.*

The refractive index alone of a transparent film is conveniently measured by the method introduced by Abelès (1950). A collimated beam of plane polarized light ($E \parallel$ plane of incidence) is incident on a surface partly covered by the film under investigation. The angle of incidence ϕ_B for which the reflectance of film and substrate are equal is determined. If the light is incident in air, then the index n of the film is simply given by

$$n = \tan \phi_B. \qquad \ldots\ldots(62)$$

This result is independent of the thickness of the film and of the optical constants of the substrate. The *sensitivity* of the method does, however, depend on these quantities. For measurements on a transparent substrate, provided that the index of the film lies within $\pm 0{\cdot}3$ of that of the substrate, an accuracy to $\pm 0{\cdot}002$ is attainable in films of suitable thickness. Since monochromatic light is used, no errors from dispersion effects arise. A particularly convenient experimental arrangement for this method is described by Traub and Osterberg (1957). The method has been applied to the study of oxide films on metals: for this measurement, a photometric detector may be needed in order to discriminate the small change in brightness around the Brewster angle ϕ_B. A suitable scanning method, in which a half-coated specimen is rotated rapidly and reflected light collected by a photomultiplier has been described by Heavens and Kelly (1959). For 50 mμ films of tantalum oxide on tantalum, an accuracy to $\pm 0{\cdot}003$ was obtained.

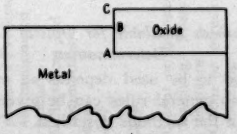


Figure 14. Schematic diagram of partly oxidized metal surface.

Measurements of the reflectance of the filmed surface in the neighbourhood of the Brewster angle may be used to check whether the film is homogeneous. The Brewster angle measurement yields an index value which may be combined with a thickness measurement using multiple-beam Fizeau fringes to enable the reflectance (on the assumption of uniformity) to be calculated using equation (13). For tantalum oxide films, significant differences between experimental and calculated values indicate a slight lack of homogeneity in the film (Heavens and

Kelly 1958). In measuring the thickness of metallic oxide films by Fizeau fringes it is necessary to measure both the steps AB and BC shown in figure 14 and to note that the fringe displacement corresponding to the step AB depends on the refractive index of the film (Kelly 1958 a).

5.2.7. *Spectrophotometric methods of thickness and index determinations.*

If white light is reflected at a film-covered surface and the reflected light dispersed, maxima and minima are observed at various wavelengths in the spectrum. The positions of such turning values are determined by the thickness and refractive index of the overlying film, although not in as simple a way as has cheerfully been supposed by some workers. The effective optical path of light in the film is made up of the optical thickness of the film together with the path equivalent of the phase change on reflection at the oxide–metal surface. This is a function of the optical constants of the film and substrate. Since all these quantities are wavelength dependent, the derivation of thickness and index from the wavelengths at turning points is a matter of some difficulty. Furthermore, the phase change on reflection depends on the direction of the plane of polarization. In studies of the oxidation of uranium, Charlesby and Polling (1955) used unpolarized light at non-normal incidence and ignored the change of phase shift with wavelength : these results must therefore be considered as very approximate. (The theory at the beginning of their paper is incorrect.) In experiments on anodic tantalum oxide, phase effects are correctly allowed for by Young (1958) : in these experiments, the refractive index was determined independently by the immersion method. The films were assumed to be homogeneous.

The magnitude of the error involved by neglect of the phase change on reflection has been estimated by Winterbottom (1946) for films on iron, copper and aluminium. The path equivalent of the phase changes are :

Fe_2O_3 on Fe : 15 mμ
Cu_2O on Cu : 22 mμ
Al_2O_3 on Al : 16 mμ

These values are appropriate to $\lambda = 546$ mμ.

5.3. *Summary of Methods available for Optical Constants and Thickness Determinations*

The choice of method to be used depends so much on the nature of the particular problem that no general rules can be given. The table opposite sets out the general features of the methods described.

§ 6. Results of Measurements of Optical Constants of Films

In this section are reviewed results serving to show the general optical behaviour found in thin films. The volume of such studies in recent years is prodigious ; detailed comment will be restricted to cases which have been reasonably thoroughly investigated. Tables of numerical values of optical constants are given at the end of the section.

Summary of Methods for Thickness and Optical Constant Measurement

Method	Quantities measured	Film		Substrate		Measurement gives			Accuracy of thickness	Remarks
		T	A	T	A	d	n	k		
Multiple-beam Fizeau fringes	Fringe displacement	Yes	Yes	Yes	Yes	Yes	—	—	±1–1·5 mμ	—
Multiple-beam Feco fringes	Fringe (wavelength) displacement	Yes	Yes	Yes	Yes	Yes	—	—	±0·5–1 mμ	Correction for phase changes
Fabry–Perot interferometer	Fringe matching. Mirror separation	Yes	Yes	Yes	Yes	Yes	—	—	±1·5 mμ	—
Michelson interferometer	Fringe matching	Yes	No	Yes	No	Yes	Yes	—	±20 mμ	—
Schopper	Amplitudes and phases of 2 reflected and transmitted beams	Yes	Yes	Yes	No	Yes	Yes	Yes	±3 mμ	Simultaneous n, k, d
Malé	Reflectances and transmittances	Yes	Yes	Yes	No	Yes	Yes	Yes	~1 mμ	Simultaneous n, k, d
Drude	Δ, ψ	Yes	No	Yes	Yes	Yes	Yes	—	~1 mμ	Valid up to 10–20 mμ
Leberknight and Lustman	Δ, ψ	Yes	Yes	Yes	Yes	Yes	Yes	Yes	—	Indirect. Curve-fitting
Vašíček	Δ, ψ	Yes	No	Yes	No	Yes	Yes	—	±1–2 mμ	Indirect. Tables given for one glass
Savart plate	Colour or intensity matching. Phase difference	Yes	No	Yes	No	Yes	Yes	—	±3 mμ	—
Zernike–Dyson	Colour or intensity matching	Yes	Yes	Yes	Yes	Yes	No	No	—	—
Stepped reflector	Phase change or fringe wavelength	Yes	No	Yes	Yes	Yes	Yes	—	±0·1 mμ or 4%	—
Schulz : k	Transmittance for several thicknesses or phase change from fringe displacement	—	Yes	Yes	No	—	—	Yes	—	Methods used together
Schulz : n	Reflectance	—	Yes	Yes	No	—	Yes	—	—	Methods used together
Abelès	Brewster angle	Yes	No	Yes	Yes	No	Yes	—	—	Very simple
Spectrophotometric	Wavelength of interference maxima and minima	Yes	(Yes)	Yes	Yes	(Yes)	(Yes)	—	—	Phase changes involved, dispersion troubles.

The optical behaviour of films is by now known to be sensitive to film structure which is in turn influenced by the conditions under which the films are formed. Assessment and comparison of much published work in this field is made difficult by the frequent absence of details of the conditions of preparation. Most of the results reported for films formed by thermal evaporation have been obtained on films produced in demountable vacuum systems at pressures in the region of 10^{-5} mm Hg. At this pressure and at the usual rates of deposition quoted, the relative numbers of depositing atoms and of residual gas atoms striking the substrate are comparable. The rate of deposition is therefore expected to influence the film properties. The ability of the deposited atoms to form into crystals will also depend on the rate of arrival of further atoms, a feature clearly shown by the electron micrographs of Sennett and Scott (1950) of silver films. Films deposited slowly are found to be much more granular than those formed rapidly.

The nature and orientation of deposited films may be influenced by the form of the substrate: under certain conditions epitaxial growth may occur. The temperature of the substrate and the effective temperature at which the film forms—a difficult quantity to measure and one over which much argument rages—also have a marked influence on the nature of the resultant film. In the results discussed below such relevant information will be included where this is available.

6.1. *Metal Films*

For most metal films a marked dependence of the optical constants on film thickness is observed. With increasing film thickness, the value of n—initially much higher than that characteristic of the bulk metal—decreases smoothly, reaching a steady value for thicknesses of the order tens of mμ. The extinction coefficient k starts from a very low value and rises, either monotonically or through a maximum, to a limiting value which may be close to the bulk value. This behaviour is interpreted in terms of the aggregated structure exhibited by the films concerned.

The dependence of optical constants on wavelength does not admit of simple generalization. Increases and decreases, with and without maxima and minima, and which are sharply dependent on film thickness, are observed. The general trend observed can be accounted for in terms of the granular structure of the film.

6.1.1. *Gold films.*

Since gold is a noble metal the properties of gold films may be expected to depend less on, for example, the effects of residual gas in the vacuum chamber and on atmospheric contamination than would those of metals susceptible to oxidation. There is a reasonable measure of agreement between different observers' results—even between those for films obtained by sputtering and by evaporation. A comprehensive bibliography of early work is given by Mayer (1950).

Dependence of optical properties on film thickness. The reflectance and transmittance, phase changes and optical constants of gold films have been studied by Rouard, Malé and Trompette (1953) and by Malé and Rouard (1953). Films of several thicknesses were prepared simultaneously by thermal evaporation through a suitably shaped rotating shield (' Talbot Disc '). It is suggested that this procedure ensures that the structures of the films are rigorously identical

although, since the rates of deposition of the various films are different, this assumption may not be justified. The rates of deposition are not given. The results of reflectance and transmittance measurements are shown in figure 15 for

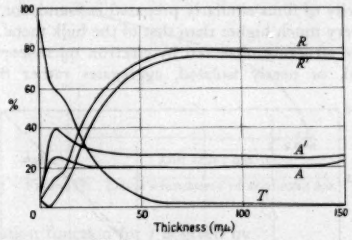


Figure 15. Percentage reflectance, transmittance and absorptance of evaporated gold films, $\lambda = 546$ mμ. (Primes indicate light incident from substrate side.)

light of wavelength 546 mμ. Similar general behaviour is shown for other wavelengths in the visible spectrum except that as shorter wavelengths are approached the limiting reflectances decrease considerably. Although the march of the reflectance–thickness curves is similar in form to that calculated from the bulk

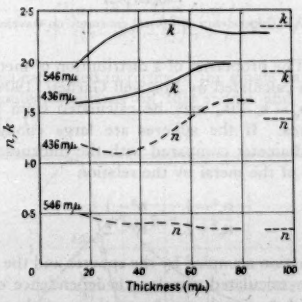


Figure 16. Optical constants of evaporated gold films. Values for annealed, thick films shown at the right-hand side.

values of the optical constants, quantitative agreement is lacking. On the assumption that the films behave as though uniform, the optical constants for the various thicknesses may be calculated. Results obtained from measurements by Malé and Rouard (1953) are shown in figure 16. There is reasonable agreement between the values of n and k for large thicknesses with the values obtained by Schulz (1957) from annealed thick films. These values may be assumed to be characteristic of the bulk metal.

Dependence of optical constants on wavelength. The observed variation with wavelength of n and k for several thicknesses of evaporated gold film is shown in figure 17, together with curves representative of the bulk material. It is found that there are marked differences from bulk behaviour for thicknesses below about 10 mμ. The resistivity of films similarly prepared is found, for thicknesses below about 10 mμ, to be very much higher than that of the bulk metal. This behaviour is consistent with the structure observed by electron microscopy which reveals a collection of isolated, or nearly isolated, aggregates rather than a continuous

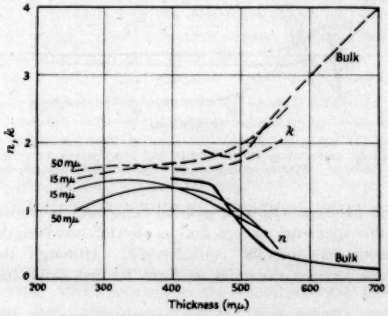


Figure 17. Dependence of optical constants on wavelength.

distribution of metal. The properties of a distribution of metal spheres embedded in a dielectric have been calculated by Maxwell Garnett (1904, 1906). The effective optical constants $\mathbf{n}_e = n_e - ik_e$ may be calculated from the constants of the constituent metal spheres. If the spheres are large enough to contain many atoms but are small in diameter compared with the thickness of the film, then $\mathbf{n}_e$ is related to the index $\mathbf{n}$ of the metal by the relation

$$\frac{\mathbf{n}_e^2-1}{\mathbf{n}_e^2+2} = q\frac{\mathbf{n}^2-1}{\mathbf{n}^2+2} \qquad \ldots\ldots(63)$$

where q is the volume fraction occupied by the spheres and the dielectric is assumed to be air. Although the calculated wavelength dependence of n_e, k_e for plausible values of q is found to show trends similar to those of the experimental results, quantitative agreement is lacking. The model is unsatisfactory for the following reasons: (i) the electron micrographs show that the metallic aggregates are not generally spherical and that their dimension normal to the film is not small compared with the average film thickness; (ii) the optical constants of such small particles will not be expected to be the same as those of the bulk since the conductivity will be considerably lowered by surface scattering. Schopper (1951), following early work by David (1939), assumes a two-dimensional distribution of spheroidal particles lying with their unique axes normal to the plane of the film. In order to obtain a fit with the experimental results it is necessary to assume that

the axial ratios of the spheroidal particles are distributed statistically about a mean value. It is convenient to work with a function f of the axial ratio b/a, calculated by David and plotted in figure 18. The axial ratios are assumed to be distributed

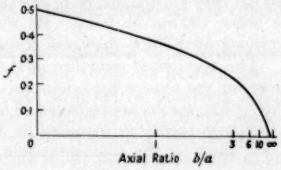


Figure 18. David's function f of axial ratio b/a.

so that the distribution function for f is given by

$$g(f) \propto f \exp(-f/\bar{f})^2. \qquad \ldots\ldots(64)$$

If d is the average film thickness and d_w the thickness deduced from the mass per unit area, then

$$\frac{d}{d_w}(n_e^2 - k_e^2 - 1) - i\frac{d}{d_w} 2n_e k_e = \bar{C} \qquad \ldots\ldots(65)$$

where

$$\bar{C} = \int \frac{[(n-ik)^2-1]\, g(f)\, df}{[(n-ik)^2-1]\, f+1} \Big/ \int g(f)\, df. \qquad \ldots\ldots(66)$$

Values of n_e, k_e are thus calculated in terms of the single parameter $\bar{f}$. The agreement between the results calculated on this model and experimental values is demonstrated in figure 19.

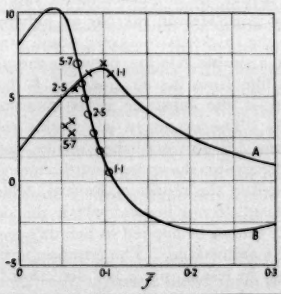


Fig. 19. Calculated and experimental results for n_e, k_e for gold films:
Curve A, $(d/d_w)\, n_e k_e$, crosses, experimental points; curve B, $(d/d_w)(k_e^3 - n_e^3 + 1)$, circles, experimental points. Thicknesses shown in millimicrons.

Optical constants of gold in the infra-red. In investigations to determine the range of wavelengths over which the Drude free electron theory is valid, measurements have been made at wavelengths out to 15 μ (Försterling and Fréedericksz 1913, Weiss 1948, Hodgson 1955, Schulz 1957). The thicknesses of films used are generally not specified but they are opaque, or nearly so, to visible light. Their thickness is thus of the order 100 mμ or more. By the time this thickness is reached the initial convulsions of the n, k–thickness curves are complete and the limiting values reached. The aggregated structure of the films may be expected to have a smaller influence on the optical behaviour for radiation of long wavelengths than for the visible. The results are collected in figure 20. The full lines are calculated from free electron theory as modified by Reuter and Sondheimer (1948) and Dingle (1953) to allow for the anomalous skin effect.

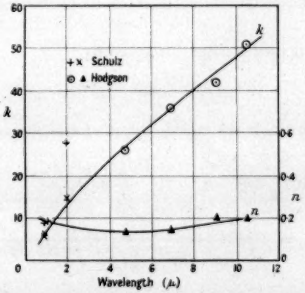


Figure 20. Optical constants of gold in the infra-red.

Homogeneity of evaporated gold films. The assumption underlying the treatments discussed above is that the properties of the film are constant throughout its thickness. Since the substrate will exercise an influence on the film structure which will diminish as the thickness increases, some gradient in properties may be expected. There is a simple relation between the absorption $\mathbf{A} = 1 - \mathbf{R} - \mathbf{T}$ for light traversing the film from the air side and $\mathbf{A}'$ for light travelling in the reverse direction. If n_s is the substrate index then for a homogeneous film $n_s \mathbf{A} = \mathbf{A}'$ (Wolter 1937). The extent to which this relation is found to hold varies considerably among different published results, indicating varying degrees of homogeneity. The departures shown by Krautkrämer's results are considerable. Fair agreement is obtained in the results of Rouard, Malé and Trompette (1953) for thicknesses up to about 15 mμ beyond which the departures become large. This test of homogeneity cannot be applied to Schulz's results since the films used for the measurement of n are opaque. The criterion applied by Schulz, that the ratio R_s^2/R_p be unity, may be less sensitive as a detector of a gradient in properties normal to the surface since both measurements are made on the same side of the film. On the other hand, for thick films such as were used by Schulz, the effects of the substrate may be felt in only a small fraction of the film thickness.

The optical behaviour of gold films deposited at oblique incidence has been studied by Reimer (1957): such films show a marked dependence of transmission on the direction of polarization relative to the evaporation direction for the film. This behaviour is associated with a structure asymmetry, clearly shown in electron micrographs of the films. Similar properties are exhibited by films of Ag, Cu and Mn.

The properties of gold films sputtered on to bismuth oxide. The differences in behaviour between films of gold produced by evaporation and that calculated for a plane, parallel-sided layer with optical constants appropriate to bulk gold are reasonably well explained in terms of the granular structure of such films. The work of Gillham, Preston and Williams (1955) on sputtered gold films deposited on sputtered bismuth oxide shows that it is possible to produce films of thickness tens of millimicrons with properties closely similar to those of bulk gold. It appears that the adhesive forces between gold and bismuth oxide are sufficiently large to overcome the tendency, produced by surface tension forces, for the film to aggregate. After suitable annealing treatment the films are found to possess electrical conductivities within about 30% of that of the bulk metal, in contrast to the behaviour shown by evaporated films. The optical constants of the films sputtered on bismuth oxide are found to be much more nearly like those of the bulk than are those of evaporated films. Electron micrographs of the sputtered films reveal continuous sheets of monocrystalline lamellae, each of lateral dimensions some ten times the mean film thickness.

6.1.2. *Silver films.*

Dependence of optical constants on thickness. There is considerably greater variation between different observers' results than in the case of gold films, probably arising from the greater reactivity of the metal, with the greater chance of contamination. There is some doubt about the figures quoted by Krautkrämer (1938), obtained by the procedure given by Murmann (1933, 1936) since this method does not give unique values of n and k. The wrong values may have been chosen.

Extensive measurements by Philip and Trompette (1957) on films up to 70 mμ in thickness and for wavelengths ranging from 302 mμ to 578 mμ show a behaviour generally similar to that shown by gold. High values of n and low values of k at small thicknesses (figure 21) tend gradually with increase of thickness to values in the neighbourhood of those recorded for the bulk metal. Analysis of the results using the Maxwell Garnett theory shows that the general features of the film behaviour are accounted for by its granular structure although, since the model is somewhat crude, the agreement is not close. Figure 21 also records the results obtained by Clegg (1952) for silver films on which the measurements were made in vacuum. The large difference between the two sets of results quoted suggests that a detailed analysis of results of this kind is better postponed until more effective control of conditions obtains.

Dependence of optical constants on wavelength. Measurements on thick silver films, annealed until the $R_s^2 = R_p$ criterion shows them to be homogeneous, were made by Schulz (1954 a) and Schulz and Tangherlini (1954) using the methods described in § 5. Whilst agreement with earlier (bulk) values of k was obtained, significantly lower values of n emerged. The measurements were made at the

glass–silver interface and were therefore less subject to uncertainties from contamination than those made on films exposed to air. For wavelengths from 1 to 3 microns the values of k are as expected from the Drude free electron theory, when the anomalous skin effect is allowed for. Departures observed in the earlier

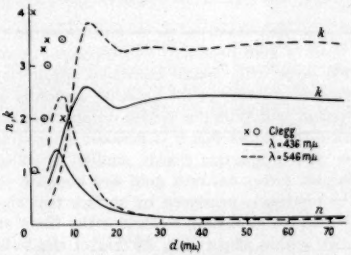


Figure 21. Optical constants of silver films.

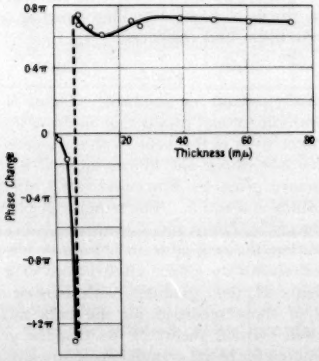


Figure 22. Phase changes on reflection at silver films.

results of Försterling and Fréedericksz (1913) are ascribed to contamination since these measurements were made on air-exposed surfaces. Recent measurements of reflectance in the region 1–13 μ by Gates, Shaw and Beaumont (1958) suffer from the same disadvantage. Since non-normal incidence and unpolarized radiation were used, the results cannot be readily interpreted. Some scatter is shown in the results for the region 1–10 μ by Beattie and Conn and Hodgson (see Schulz 1957), reflecting the difficulties attendant in this spectral region, but the values of k are substantially as expected on the Drude theory.

Homogeneity of evaporated silver films. Extensive measurements on the phase changes on reflection at films of silver (Avery 1950, Faust 1950, Clegg 1952, Philip 1956, Dumontet, Perrot and Tortosa 1957, Lisitsa and Tsvelykh 1957) reveal a highly complex variation of this quantity with film thickness and with wavelength. For some wavelengths a phase jump of 2π (figure 22) is observed for a thickness generally below 10 mμ. At other wavelengths no such jump is found. This behaviour cannot be interpreted in terms of a single homogeneous film and measurements by Dumontet *et al.* suggest that a transition layer at the surface could be the cause of the anomalies. The results of phase change measurements made on silver surfaces exposed to air are of considerable importance to interferometry : it is unlikely that a detailed study of the adventitious contamination which gives rise to these anomalies is worth while.

A summary of results of optical constant measurements on thick silver films at various wavelengths is given in Appendix II.

6.1.3. *Other metals.*

Alkali metals. Measurements of the optical constants of (unspecified) thick films of potassium, sodium, rubidium and caesium have been made by Ives and Briggs (1936, 1937 a, b) for wavelengths from 200 mμ to 650 mμ. The films were condensed on quartz and sealed against atmospheric contamination by

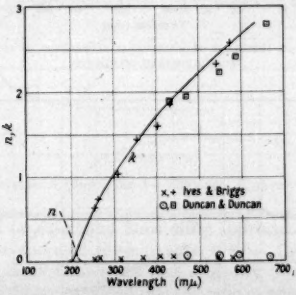


Figure 23. Optical constants of sodium. Curves calculated from Kronig theory.

paraffin wax, measurements being made at the quartz–metal surface. The results are found to agree reasonably well with those calculated on the basis of Bloch's theory of metallic conduction, with the assumption that collisions between electrons and the atoms of the lattice may be neglected (Kronig 1931). Dispersion curves for sodium are shown in figure 23 together with early measurements of Duncan and Duncan (1913). Collected values of n and k for the alkali metals are given in Appendix II.

Aluminium. In the far ultra-violet region of the spectrum the reflectance of aluminium is among the highest known. On account of their importance in spectroscopic work, the properties of Al films have been carefully studied in this

region (Hass, Hunter and Tousey 1956, 1957, Gerharz 1958). The speed of evaporation has a pronounced effect on the reflectance, particularly in the region $\lambda = 160$ mμ to 180 mμ, the highest reflectance being obtained for rates of deposition in excess of 30 mμ sec^{-1}. No marked dependence on pressure in the neighbourhood 10^{-5} to 10^{-4} mm Hg is observed. The dependence of reflectance on film thickness for various wavelengths is shown in figure 24. A fall in reflectance is observed on ageing and is shown to be a consequence of the formation of aluminium

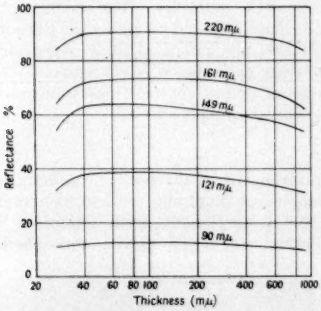


Figure 24. Dependence on thickness of ultra-violet reflectance of aluminium films. Wavelengths marked on curves.

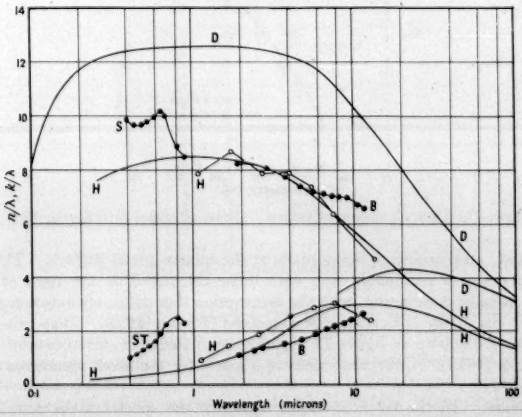


Figure 25. Calculated and experimental values for n/λ and k/λ for aluminium : B, Beattie 1955 ; H, Hodgson 1955 ; ST, Schulz and Tangherlini 1954 ; S, Schulz 1954 a ; D, Drude theory.

oxide on the surface. The optical constants of aluminium and aluminium oxide in this region, determined by reflection measurements, are given in table 1.

Table 1. Optical Constants of Al and Al_2O_3 in the Ultra-violet

Wavelength (mμ)		90	122	220
Aluminium	n	0·55	0·37	0·14
	k	0·60	0·94	2·35
Aluminium oxide	n	1·30	1·75	1·85
	k	0·87	0·72	0·00

In the visible region, early results of O'Bryan (1936) are substantially confirmed by Schulz and Tangherlini (1954) and Schulz (1954 a). Measurements in the near infra-red by Hodgson (1955) and Beattie (1955) show that, although the results agree with predictions of the Drude theory on the basis of approximately 1·3 electrons per atom (figure 25), this assumption leads to incorrect behaviour at very long (100 μ) wavelengths, at which the classical skin effect would be expected to apply.

Copper. The variation with film thickness of the optical constants of copper films (figure 26) is similar to that shown by silver and gold. The limiting values

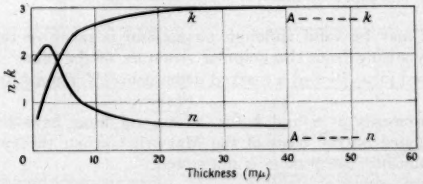


Figure 26. Optical constants of copper for $\lambda = 586$ mμ. A, Avery's results on thick films.

to which n and k tend for sputtered films differ appreciably from those obtained for thick (40–50 mμ) evaporated films (Avery 1950). Results of measurements by Schulz (1954 b) on thick films show that a fit may be obtained at long wavelengths with the Drude theory provided that the ratio of effective to normal electron mass is taken as 1·45. This value agrees with that obtained from electronic specific heat determinations. Values of m^*/m deduced from the measurements of Beattie and Conn (1955) in the near infra-red yield a considerably higher value, suggesting that bound electrons may be playing a part in the absorption. There is some considerable difference between Beattie and Conn's results and those of Hodgson (1955), presumably due to some difference in experimental conditions.

Platinum and palladium. Early measurements on sputtered films of these metals show the usual trends of n and k with film thickness. More recent work of Malé and Trompette (1957) on Pd films formed by thermal evaporation (pressure 10^{-5} mm Hg, rate of deposition 10 mμ min^{-1}) are interpreted in terms of film granularity. From the mass thickness d_m and the figure deduced optically d_o, the *coefficient de remplissage* $q = d_m/d_o$ is determined. Values of q obtained for

several different wavelengths (302 to 546 mμ) are found to agree, thus validating the procedure adopted for this metal. The optical constants are found to increase monotonically with thickness for all wavelengths. The dependence of q on the thickness d_0 is shown in figure 27 and reveals a very large extent of voids in the thinner films. In analysing the above results in terms of dispersion theory Grard (1957 a) concludes that there are no resonant frequencies over the wavelength range examined, in contrast to those observed for gold and silver (Grard 1957 b, c).

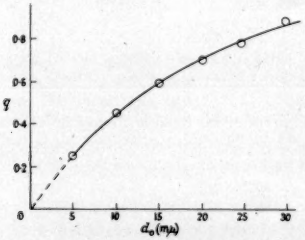


Figure 27. Dependence of q on thickness for Pd films.

This conclusion may be valid although no account is taken of the variation of optical constants arising from the granular structure of the films. In the thickness range covered (5 to 30 mμ) a marked dependence of n and k from this cause is observed.

Tin. Measurements of n and k for 2·5 mμ tin films have been compared with those calculated on the basis of the Maxwell Garnett theory. Reasonable agreement is obtained: the q-value is not stated.

Table 2. Optical Constants of Tin Films. Thickness 2·5 mμ

Wavelength	356 mμ		546 mμ	
	n	k	n	k
Experimental	2·3	1·0	3·0	0·5
Calculated	2·2	1·5	2·8	0·3
Bulk metal	—	—	1·0	4·2

Measurements by Hodgson (1955) on thick tin films in the infra-red show departures from the free electron behaviour at wavelengths below about 5 μ. No account is taken of the highly granular structure which is observed even in very thick films and this might well account for the departures from theory which treat the film as a continuum.

Lead. Measurements of the optical constants of lead films *in vacuo* and in air (Trompette 1959) show the usual trends suggestive of a granular structure, the values of q ranging from 0·49 to 0·82 for films of thickness up to 35 mμ. Whilst the optical behaviour of these films for visible wavelengths changes but little on exposure to air, the ultra-violet absorption rises considerably. This is rather strange in view of the fact that the conductivity of the films decreases for

thicknesses below 65 mμ. The reason for this is not clear. The effect of oxidation on the optical properties is found to occur practically instantaneously suggesting that a completely protective oxide film forms.

Tantalum, tungsten, zirconium and silicon. These metals prove of particular interest in the far ultra-violet where their reflecting properties approach that of aluminium. Measurements by Fabre and Romand (1956) on films formed by evaporation (15 minutes' evaporation for a film transmitting approximately 1% of visible light) are shown in figure 28. The curve for tungsten refers to a freshly

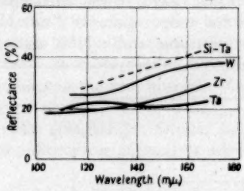


Figure 28.
Reflectance of films of Ta, W and Zr. Broken line: Si film contaminated with tantalum.

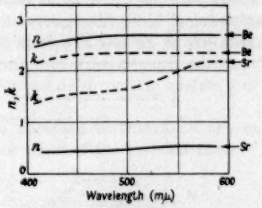


Figure 29.
Optical constants of Be and Sr.

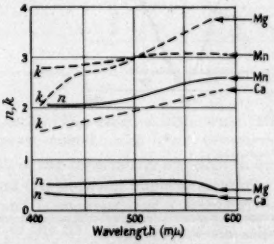


Figure 30.
Optical constants of Ca, Mn and Mg.

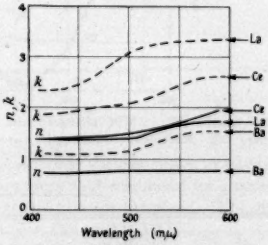


Figure 31.
Optical constants of Ba, Ce and La.

deposited film: on exposure to air a considerable fall in reflectance results. Although the reflectance of Si films in this region is about 10%, that of films formed by high-temperature evaporation from a Ta boat is very high (figure 28) (Kandare and Fabre 1956).

Mg, Be, Ca, Ba, Sr, La, Mn and Ce. O'Bryan (1936) has measured the optical constants of these metals for the principal lines of the mercury spectrum in the visible: the results are collected in figures 29–31. The films were all produced by thermal evaporation.

6.2. *Dielectric Films*

The optical behaviour of dielectric films is in some ways simpler than that of metal films. Whereas metal films show near-dielectric properties in very small

thicknesses, on account of the effects of granularity, in a dielectric film this feature simply results in a decreased refractive index. The effects of granularity in structure and of inhomogeneity are however detectable, as illustrated by the results obtained for cryolite and calcium fluoride discussed below. Films produced by thermal evaporation will be discussed first followed by those produced chemically or electrolytically.

6.2.1. *Films formed by thermal evaporation.*

Calcium fluoride. A detailed study by Bousquet (1956) of the reflectance and transmittance of CaF_2 films, deposited by thermal evaporation (10^{-5} mm Hg) on glass and quartz, reveals that such films are anisotropic and exhibit a *couche de passage* at the film–air surface. For a uniform film whose index is lower than

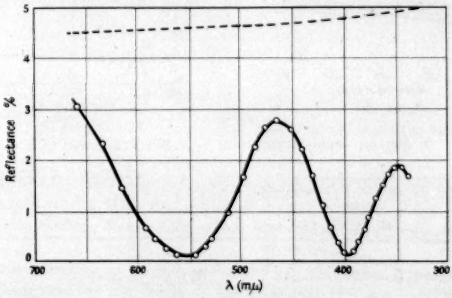


Figure 32. Reflectance–wavelength curve for evaporated CaF_2 films. Broken curve: reflectance of bare substrate.

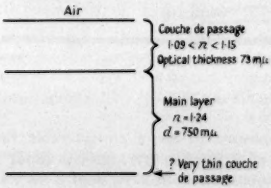


Figure 33.

that of the substrate and of optical thickness $k\lambda/2$ the reflectance should exhibit maxima equal to the reflectance of the bare substrate. The spectrophotometric curve for a film of optical thickness 690 mμ on glass is shown in figure 32. The maxima lie well below the values for the bare substrate. The transmittance curve is found to be complementary, within the accuracy of the measurements (to $\pm$ 0·3%), showing that the departure from the behaviour expected of a uniform film cannot be ascribed to absorption. Figure 33 shows the form of layer which would yield

the reflectance curve of figure 32. When monochromatic light is totally reflected at a glass–CaF_2 interface, fringes are observed in the neighbourhood of the critical angle, consistent with the diffusion of light by the film at angles other than the apparent angle. A study of the contrast observed in the fringe system suggests that these films are anisotropic, a conclusion amply confirmed by phase change measurements at angles near the Brewster angle (Bousquet and Delcourt 1957). The magnitude of the birefringence is found to depend on the angle of arrival of the condensing atoms (Sinelnikov, Shkliarevsky and Vlasenko 1957) the maximum birefringence observed being about 0·03. Films of fluorite deposited at oblique incidence on a rotating surface are found to show marked optical activity (Young and Kowal 1959), the rotation for films 18 μ thick ranging up to the order 2°. Films deposited obliquely on to a fixed substrate show birefringence with an optic axis inclined to the film normal. On a film formed on a rotating surface a helical anisotropy results.

Cryolite (Na_3AlF_6). Similar behaviour is observed to that for CaF_2 except that the reflectances at maxima are higher than those of the bare substrate. The

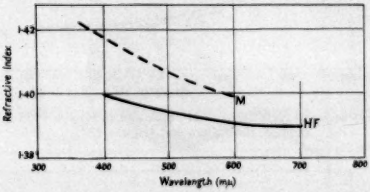


Figure 34. Dispersion of MgF_2 films. HF, Hall and Ferguson 1955 a; M, Morita 1956.

march of the reflectance–wavelength curves measured at various angles of incidence around the Brewster angle is markedly different from that calculated for a uniform film and is suggestive of a *couche de passage* at the film–air interface. The mean value of the refractive index of the whole film ($\lambda = 486$ mμ) is 1·375.

Magnesium fluoride. Since this material is widely used as an anti-reflecting coating on glass, its properties have been extensively studied (Bannon 1946, Rood 1949, Abelès 1950, Schulz and Scheibner 1950, Heavens and Smith 1957, Sanford 1958). Reflectance measurements by Bousquet (1956) indicate that, unlike CaF_2 and cryolite, the departures from homogeneous behaviour are relatively small. Since the values obtained for the refractive index lie close to (in fact, *above*) those quoted for the mineral sellaite, the films probably contain few voids.

Refractive index measurements by Hall and Ferguson (1955 a) and by Morita (1956) are shown in figure 34 for films of 200 to 800 mμ thickness. Electron micrographs of MgF_2 films show them to be exceedingly fine-grained (Morita 1957). Loss by scattering in films formed on glass is found to be less than 1%: on heated quartz, a loss of several per cent is observed, arising from the formation of cracks in the film. MgF_2 films are found to be in a state of high tensile stress which increases with thickness and may cause disruption of the film at thicknesses beyond 2 μ. In the ultra-violet region of the spectrum, the refractive index is

found to depend more markedly than in the visible on the film thickness and rate of deposition. Marked inhomogeneity is detected (Hall 1957), especially in rapidly deposited layers of thickness a few tens of millimicrons. Absorption in the films—negligible in the visible—rises steeply for wavelengths below 230 mμ ($k \sim 0{\cdot}005$ at $\lambda = 210$ mμ).

Lithium fluoride. In studies of thick (2–18 μ) films of LiF Schulz (1949) finds that the proportion of solid LiF in the films rises from 70 to 82%. The voids are mainly accessible to a filling liquid although as the thickness increases over the above range the proportion of inaccessible voids rises to about 3%.

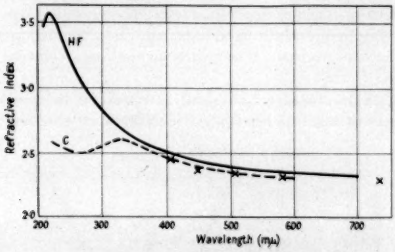


Figure 35. Refractive index of ZnS films. HF, Hall and Ferguson 1955 a ; C, Coogan 1957 ; crosses, Kuwabara and Isiguro 1952.

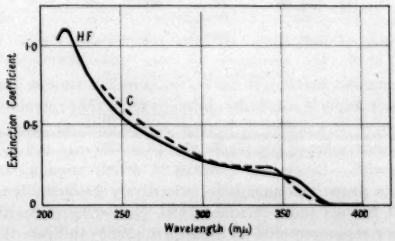


Figure 36. Extinction coefficient of ZnS films. HF, Hall and Ferguson 1955 a ; C, Coogan 1957.

Zinc sulphide. The popularity of this material as a high-index component in multilayers, before being largely ousted by cerium dioxide, has resulted in many studies of its properties (Hammer 1943, Shalimova 1951, Hall and Ferguson 1955 a, b, Hall 1956, Coogan 1957). Over the region 400 mμ to 700 mμ there is reasonable agreement between different results, showing a decrease in index from 2·5 at 400 mμ to 2·3 at 700 mμ. Most of the values obtained for films are slightly below ($\sim 0{\cdot}1$) those for the bulk crystal. Below 400 mμ there are large differences, both between different observers' film results and between film and bulk values

(figure 35). Hall records a steady increase in n as λ falls to 225 mμ, at which wavelength $n = 3{\cdot}6$. Coogan's results show a maximum (2·6) at 350 mμ and a minimum (2·5) at 290 mμ. Fair agreement is shown in the results for the extinction coefficient (figure 36). It is suggested that the enhanced absorption in the neighbourhood of 330 mμ may indicate an exciton band. From 2 μ to 14 μ the refractive index of ZnS films falls steadily from 2·25 to 2·15 : the extinction coefficient is less than approximately 0·001 in this region.

Huldt and Staflin (1959) obtain a significantly higher index ($\sim$2·37 at $\lambda = 1{\cdot}5$ μ) for films prepared with residual air in the evaporation chamber than for those produced with residual nitrogen ($n \sim 2{\cdot}28$). Similar behaviour is shown by germanium films.

X-ray studies indicate that ZnS films deposited on glass substrates at room temperature are mainly amorphous.

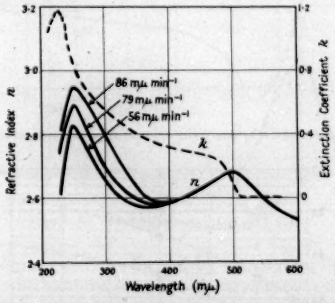


Figure 37. Dependence of refractive index of CdS films on rate of deposition. Full lines, refractive index ; broken line, extinction coefficient.

Cadmium sulphide. The optical behaviour of CdS films is found to depend on the rate of deposition, low rates favouring low optical absorption and low refractive index (figure 37). Considerable care is needed if spurious high absorption, due to dissociation, is to be avoided. Hall and Ferguson evaporate the material from a filament in contact with the charge. The filament temperature must not exceed 1025°c. In the region 1–4 μ, the extinction coefficient is a minimum, ranging from approximately 0·01 for a deposition rate of 40 mμ min^{-1} to approximately 0·05 at 200 mμ min^{-1}. (It seems likely that this variation will be due to dissociation.) From 2 μ to 14 μ, the refractive index falls slightly from 2·30 to 2·25. Beyond 7 μ, k is no longer influenced by the rate of deposition. The films show the cubic structure of CdS.

Cerium dioxide. The high refractive index and robust mechanical properties of CeO_2 films make them very suitable for high-reflecting multilayers (§ 8.2) in place of ZnS. The conditions necessary for obtaining reproducible, stable properties have been investigated by Hass, Ramsey and Thun (1958). The crystallite size (figure 38) and refractive index (figure 39) depend markedly on the

substrate temperature. Before evaporation, CeO_2 powder is baked and sintered in air at 1100° c: furthermore the evaporation charge must be pre-heated to the evaporation temperature before deposition is allowed to begin. CeO_2 films show considerable absorption ($R = 0{\cdot}41$, $T = 0{\cdot}24$, $A = 0{\cdot}35$) at $\lambda = 320\,m\mu$. At 500 mμ and above the absorption is negligibly small ($< 0{\cdot}1\%$ in $\lambda/4$ layers).

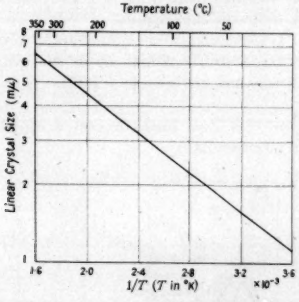


Figure 38. Crystallite size of CeO_2 films as function of substrate temperature during deposition.

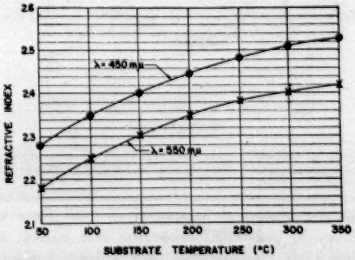


Figure 39. Dependence of refractive index of CeO_2 films on substrate temperature.

Antimony sulphide. The few results obtained for this material show such wide differences that only brief mention is made here. Schopper (1952 b) obtains $n = 2{\cdot}5$, $k = 0{\cdot}01$ for $\lambda = 640$ mμ and $n = 2{\cdot}6$, $k = 0{\cdot}29$ for $\lambda = 436$ mμ. These are significantly lower than values reported by Kunze (1958) who finds a maximum (3·4) at 680 mμ falling to 2·7 at 500 mμ and 1·5 μ. These are in turn lower than the results of Billings and Hyman (1947): $n = 3{\cdot}0$ at $\lambda = 589$ mμ. There seems some doubt over the optical constants of the mineral stibnite although the values quoted are very much higher than those obtained on films (Hutchinson 1907, Bailly 1938).

Silicon monoxide (Hass and Salzburg 1954). In order to form films of SiO by evaporation, a high rate of deposition (> 1 mμ sec^{-1}) is needed at a pressure less than

10^{-5} mm Hg ; the source temperature should be approximately 1200°c. Unless these conditions are met there is a tendency for the higher oxide to form. (Since the absorption of SiO_2 in the ultra-violet is lower than that of SiO, the oxidized film is more suitable as a protective film in this region.) The optical constants of (true) SiO films of thickness 30–2000 mμ are shown in figure 40. The absorption peak at 10 μ results in a decrease in reflectance of an aluminium film with protective SiO coating if too large a thickness of SiO is used. For a protective film 125 mμ thick the decrease amounts to only 1–2% (the transmission of a rocksalt plate is, however, reduced by about 20%).

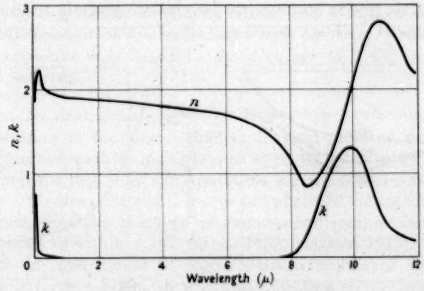


Figure 40. Optical constants of SiO films.

An interesting property of films formed by evaporating SiO is the wide range of refractive indices which can be obtained by varying the evaporation conditions. In the wavelength range 500 to 700 mμ indices from 1·55 to 2·00 may be obtained. The mechanical properties of the films are satisfactory.

6.2.2. *Chemically and electrolytically deposited films.*

Titanium dioxide. Films formed by the direct evaporation of TiO_2 are found to be heavily absorbing, indicating that some decomposition takes place. Transparent films may be produced either by allowing fuming $TiCl_4$ to decompose on heated glass (Banning 1947) or by thermal oxidation of titanium films formed by evaporation (Hass 1952). Films formed by oxidizing Ti at 400°c for several hours are found to consist of the rutile form of TiO_2. The measured dispersion of the films follows closely that of bulk rutile, the film values being consistently about 0·05 below the bulk figures. Over the wavelength range 400 to 2000 mμ the absorption of $\lambda/4$ films is less than 0·5%. The refractive index falls steadily from 3·00 at 400 mμ to 2·50 at 2000 mμ. Films formed by the decomposition of $TiCl_4$ vapour on glass at temperatures below 280°c are found to be amorphous : at higher substrate temperatures the anatase form of TiO_2 is produced. The refractive index of these films ranges from 2·4 (450 mμ) to 2·25 (700 mμ). They are rougher and mechanically less durable than those made by oxidizing metallic Ti films.

Aluminium oxide. Transparent, non-porous films of aluminium oxide can be prepared (Hass 1949) by anodizing in a solution of ammonium tartrate. The films grow to a limiting thickness determined by the applied voltage. This thickness is approximately 1·3 mμ volt^{-1}, depending to some extent on the state of the aluminium surface before anodizing (van Geel and Schelen 1957). The exact nature of the films is uncertain: since they are amorphous, identification by x-ray means is not possible. The observed refractive index ($\sim 1{\cdot}6$ for $\lambda = 589$ mμ) lies well below that of bulk Al_2O_3 (1·76); also the density is found to be 3·1 g cm^{-3} whereas that of the bulk is 3·42 g cm^{-3}. Refractive index measurements by Weisskirchner (1951) reveal an increase with anodizing voltage for measurements at 589 mμ. Harris (1955) observes the reverse tendency in measurements made out to 15 μ (figure 41). The extinction coefficient shows an absorption band near

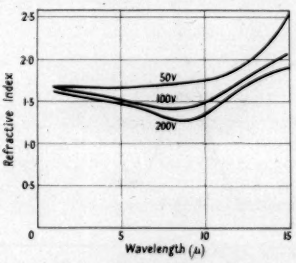


Figure 41.

Dispersion of anodic aluminium oxide films. Anodizing voltage shown on curves.

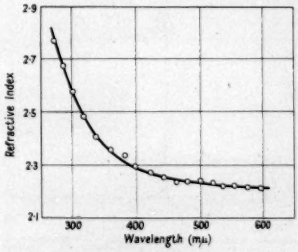


Figure 42.

Dispersion of anodic Ta_2O_5 films.

15 μ; values of k are scarcely affected by the anodizing voltage. The refractive index results suggest that there may be a dependence of film composition or structure on thickness. That there is some uncertainty over the structure is shown by measurements by Kelly (1958 b) of the volume ratio of the oxide: this is found to depend on the age of the electrolysing solution.

Tantalum oxide. The most recent results on the refractive index of anodic tantalum oxide (Young 1958) are shown in figure 42. The index at 590 mμ was determined by immersion: the indices relative to this value were found from spectrophotometric curves. In so far as the correct procedure is adopted in allowing for phase changes on reflection at the interfaces, these results are correct for the average index of the films. There is some evidence (Heavens and Kelly 1958) that these films show a gradient of index throughout their thickness.

Reactively sputtered oxides. Films of certain metal oxides may be produced by sputtering the metal in a cathodic discharge in the presence of oxygen—a process known as reactive sputtering. Such films are generally semiconducting with a reasonably high (from ~ 100 ohms per square) conductivity: many are non-absorbing in some parts of the visible spectrum. Details of the methods of preparation and of the optical and electrical properties of iron oxide, indium oxide, tin oxide and cadmium oxide are given by Holland and Siddall (1953).

§7. Methods of Controlling Film Thickness

In many of the applications of thin films to optical systems discussed below it is necessary to deposit films of accurately determined thickness. Several methods for achieving this are available. The nature of the problem is such as to require methods which can be employed during the deposition of the film: both polarimetric and spectrophotometric methods can be used.

7.1 *Semi-reflecting Metal Films*

In certain types of filter, partly reflecting metal films are used although in many instances these would now be replaced by multiple dielectric layers. The thickness of the metal layers is generally not highly critical and it suffices to monitor their thickness by measurement of the transmitted light intensity. For opaque films, resistance measurements may be used. An accuracy to 3–4% is easily attained and is generally sufficient.

7.2 *Dielectric Layers*

Visual observation of the colour changes in light reflected from a dielectric layer during deposition enables multiple quarter-wave layers to be deposited with an accuracy to perhaps 10%: for some purposes, e.g. single anti-reflecting coatings, this is sufficient. If a monochromatic source and photocell are used, preferably with a recording meter, then the positions of maxima and minima in the reflectance curve can be located to within $\pm 2\%$ (Jarrett 1952, Heavens 1955). This accuracy is adequate for the preparation of high-reflecting multilayer stacks. For the production of interference filters, in which the position of the passband depends critically on the layer thickness, a higher accuracy is needed. The method introduced by Giacomo and Jacquinot (1952) entails illuminating the system with a beam whose wavelength scans over a small range with frequency f. If the light passes through a film of thickness $m\lambda/4$, where λ is the mean of the scanning wavelength range, then the photocell output will contain no signal of frequency f. If λ corresponds to a point of inflection on the transmission curve of the film a zero at frequency $2f$ is obtained. This method has been further developed by Lissberger and Ring (1955) and yields an accuracy to ± 1 mμ in the position of the passband of a second-order interference filter (§8.3) and to $\pm 1{\cdot}5$ mμ in that of a first-order filter. A simple experimental arrangement using colour filters is described by Traub (1956) in which duplication of a given film of the correct thickness is achieved to within a small fraction of one per cent.

Polarimetric methods have been employed by Hermansen (1955) and by van Heel and Walther (1958). Hermansen reflects plane polarized light at an angle of 75° and measures the differential phase change $\Delta = \delta_p - \delta_s$. For suitably chosen wavelengths an accuracy to ± 1 mμ is obtained. The normal incidence method of van Heel and Walther (§5.2) using a Savart plate is experimentally somewhat more convenient: the accuracy attainable is to ± 2 mμ.

7.3 *Piezoelectric Crystal Method*

In a recent and novel method by Lostis (1959) the change in frequency of an oscillating quartz crystal is measured when a film is deposited on its surface. The frequency difference between that of the test crystal and of a standard is measured

using an ordinary counter-type frequency meter. For a resonant frequency of 5 Mc/s the frequency change can be easily measured to within ± 1 c/s. For a crystal of mass M_0 of resonant frequency f_0 the frequency change Δf for a film of thickness Δe and density ρ is given by

$$\frac{\Delta f}{f_0} = \rho \frac{\Delta e}{M_0}. \qquad \ldots\ldots(67)$$

For an aluminium film for example a thickness of 1 mμ produces a change of 3 c/s indicating that the method is capable of an accuracy to about $\pm 0{\cdot}3$ mμ.

§ 8. Some Applications of Thin Films in Optics

The number and diversity of applications of thin film techniques is prodigious and is increasing rapidly. No attempt is made to be comprehensive ; details are given of results which serve to illustrate the type of system which can be achieved with existing materials and methods.

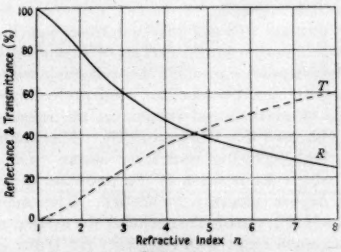


Figure 43.

Reflectance and transmittance of parallel-sided (thick) slab for different refractive indices.

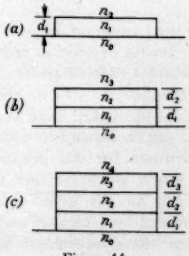


Figure 44.

Typical anti-reflection systems. $n_1 d_1 = n_2 d_2 = n_3 d_3 = \lambda/4$.

8.1 *Anti-reflecting Systems*

In some circumstances the loss of light by reflection at the surfaces of the optical elements of a system can be serious. Figure 43 shows the transmittance of a plane parallel-sided slab of non-absorbing material as a function of refractive index : indices as high as 4–5 are commonly met in optical systems in the near infra-red.

For reducing reflection in the visible spectrum magnesium fluoride offers a satisfactory compromise between anti-reflecting properties and durability. The colour effects likely to result when many elements of a system are coated with MgF_2 layers have been investigated by Murray (1956). Double and triple layer anti-reflecting systems may be designed to reduce reflectance to zero at two and three wavelengths (Lockhart and King 1947, Turner 1950). For the systems shown in figure 44 the relations to be satisfied for zero reflection are as follows :

(*a*) $n_1^2 = n_0 n_2$ .....(68)

(*b*) $n_2^2 n_0 = n_1^2 n_3$ (single zero) .....(69)

(*c*) $n_1 n_2 = n_0 n_3$ (two zeroes) .....(70)

(*d*) $n_2^2 = n_1 n_3 = n_0 n_4$ (three zeroes). .....(71)

The positions of the multiple zeroes with respect to λ_0 are determined by the values of the indices chosen (equations (69)–(71) determine only the *ratios* of the indices). In particular, for the three-layer system, the outer zeroes lie symmetrically about λ_0 if $n_1^2 = n_0 n_2$ and $n_3^2 = n_2 n_4$: this system gives effective reduction of reflectance over a wavelength range three times that for a single film.

If an electronic computer is available approximate methods for obtaining a desired reflectance curve may be used similar to that described in § 8.2.3 for a broad-band reflecting system. The curves for a given set of indices and thicknesses may be computed directly from equation (38). Approximate expressions for 2-, 3- and 4-layer anti-reflecting systems have been given in closed form by Thelen (1956) and by Kard (1957 b). For systems containing no high-index layers, simple graphical methods in which multiple reflections are ignored serve to indicate the general behaviour.

Systems used in the near infra-red.

For plates required to transmit well over a fairly narrow ($\sim \lambda/3$) region, a single quarter-wave film suffices. An increase of 5–10% in the transmittance of quartz, flint glass, MgO and sapphire plates (Jeness 1956) by the use of $\lambda/4$ MgF_2 and LiF layers is reported. In figure 45 are compared the transmission of a silicon plate (1·5 mm) with and without anti-reflecting layers of SiO, evaporated to give a refractive index of 1·90 (Cox and Hass 1958). For longer wavelengths ZnS has been used. The effectiveness of a thin InSb slab as a cut-on filter is greatly improved (Smith and Moss 1958) by anti-reflecting films of $PbCl_2$ (figure 46).

For plates which are required to operate over wider ranges two- and three-layer coatings have been used. Reflection and transmission curves for glass with (*a*) a single MgF_2 layer and (*b*) a $SiO + MgF_2$ pair are shown in figure 47 (Cox, Hass and Rowntree 1954). Arsenic trisulphide glass may be effectively bloomed by $\lambda/4$ cryolite + $\lambda/4$ tungsten oxide layers resulting in an improved transmission from 1·6 to 8 μ (apart from an absorption band at 3 μ, possibly due to water in the films). Triple-layer blooming of Ge, with stibnite (2·8), tin oxide (1·9) and thorium fluoride (1·5) yields more than 90% transmission from 3·5 μ to 10 μ, in contrast to 45% for the uncoated plate (figure 48) (Hass and Turner 1957).

8.2 *High-reflecting Systems*

8.2.1. *Protected metal mirrors.*

By the use of suitable thicknesses of transparent protective films on metal mirrors the reflectance may be increased over part of the spectrum. The use of SiO as a protection for Al has been mentioned above: it should be noted that whereas the absorption of a 125 mμ SiO film on a transparent window (e.g. NaCl) is considerable at some wavelengths, the effect on reflection from a metal mirror is negligible. This follows from the fact that a node forms at the metal surface: for wavelengths much greater than the optical film thickness the energy absorbed under this condition is very small. Protection and enhancement of reflection can be effected by quarter-wave pairs of $Al_2O_3 + TiO_2$ or of $MgF_2 + CeO_2$. A significant increase in reflectance in the visible is found for aluminium mirrors coated in this way (Hass 1955).

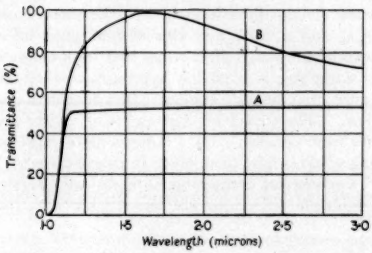


Figure 45. Si plate bloomed with SiO films. A, uncoated plate; B, coated plate.

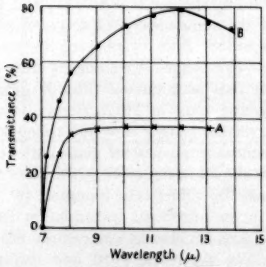


Figure 46.
InSb slab bloomed with $PbCl_2$ films. A, uncoated plate; B, coated plate.

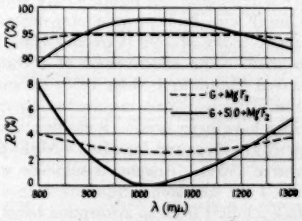


Figure 47.
Glass bloomed with one- and two-layer anti-reflecting coatings (near infra-red).

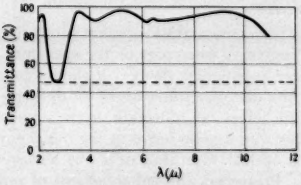


Figure 48. Three-layer anti-reflecting system on germanium (Hass and Turner 1957). Broken line, uncoated plate.

8.2.2. *Reflection filters.*

Selective reflection over a limited wavelength region may be effected by depositing interference combinations of dielectric and metal films on an opaque mirror (Hadley and Dennison 1947, 1948, Turner and Hopkinson 1953). By the use of semiconducting films which possess high absorption in the visible but are transparent in the infra-red, mirror combinations can be made which are practically non-reflecting in the visible but highly reflecting beyond about 0·8 μ. The position at which the cut-on wavelength occurs may be adjusted partly by choice of

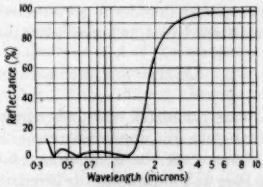


Figure 49. Reflection characteristic of dark mirror Al–SiO–Al–SiO (Hass, Schroeder and Turner 1956).

base metal and partly by the thickness of the layers. Figure 49 illustrates the type of reflectance curve which can be obtained in this way (Hass, Schroeder and Turner 1956).

For a system required to transmit visible light and reflect infra-red, stacks of non-absorbing layers may be used. Suitable designs are discussed by Epstein (1955).

8.2.3. *All-dielectric reflection systems.*

High reflectance around λ_0 is easily attained by the use of multilayer stacks of alternate $\lambda_0/4$ high- and low-index layers. Table 3 gives the calculated maximum reflectances attainable at normal incidence for three typical systems (eqn (34)).

Table 3. Calculated Reflectance at Maximum of Quarter-wave Stacks

System	CeO_2 + cryolite λ = 550 mμ	$Sb_2S_3 + CaF_2$ λ = 1 μ	Ge + cryolite λ = 2 μ
DH	0·35	0·447	0·69
DHLH	0·73	0·835	0·96
DHLHLH	0·91	0·900	$0{\cdot}99_8$
DHLHLHLH	0·97	0·977	—
	n_D = 1·52	n_D = 1·45	n_D = 1·45

D = dielectric substrate; H = $\lambda/4$ high-index layer; L = $\lambda/4$ low-index layer.

The low absorption of multilayer dielectric stacks compared with that of metallic reflecting films is of considerable advantage in interferometric applications. Nine-layer ZnS + cryolite systems, with a reflectance of 97–98% in the visible, absorb less than 1%. (The loss is probably by scattering rather than by true absorption.) The application of quarter-wave multilayers to the Fabry–Perot etalon (Steudel 1957, Giacomo 1958, Ring, Beer and Hewison 1958) shows that the optical

performance of an etalon employing optical flats is limited by the irregularities of the flats. The use of a mica spacer is suggested: this makes use of the fact that certain grades of mica cleave to form parallel-sided sheets which are smooth to within one lattice spacing (~ 2 mμ).

Broad-band reflecting system. The bandwidth over which high reflectance is shown may be broadened by the use of staggered layer thicknesses. Baumeister and Stone (1956) give a simple procedure for computing the characteristics of a coating to give a broad reflectance band over the visible spectrum. The computation is performed first for a system of n layers of optical thicknesses $\lambda_i/4$ ($i = 1, 2, \ldots, n$). If we assume that, in order to get a better fit with a given reflectance curve, the thicknesses are altered by $\delta\lambda_i$, we may write for the reflectance change at any wavelength:

$$\delta R(\lambda) = \sum_n A_i(\lambda)\,\delta\lambda_i \qquad \ldots\ldots(72)$$

to first order. The $A_i(\lambda)$ are calculated, using the computer, from equation (38). $A_i(\lambda) \equiv (\partial R/\partial\lambda)_i$. From the given and approximate curves δR is read off at n wavelengths; the computer then solves the n simultaneous equations for $\delta\lambda_i$. The process is repeated for as long as the computer is hired. The resultant reflectance

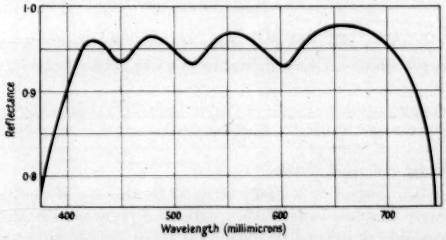


Figure 50. Broad-band reflecting multilayer systems (Baumeister and Stone 1956).

curve for a 15-layer system of $\lambda_i/4$ layers of ZnS and cryolite is shown in figure 50. The λ_i values used are given in table 4.

Table 4. Broad-band Multilayer Design: values of λ_i

Layer No.	λ_i (mμ)	Layer No.	λ_i (mμ)
1 (ZnS)	690·8	9	520·5
2 (cryl)	690·8	10	463·7
3	690·8	11	463·7
4	666·7	12	434·8
5	575·7	13	414·0
6	701·3	14	414·0
7	626·2	15	414·0
8	517·0		

Such a system exhibits a very high dispersion of phase change on reflection (Baumeister and Jenkins 1957); when used in an interference filter very sharp

bands result, occurring at wavelengths which are less critically dependent on the spacer thickness than in the simple all-dielectric filter.

The use of multilayers extends the limit of Fabry–Perot interferometry into the infra-red. Combination of a grating instrument with an etalon enables very high resolution to be attained. With multilayers of reflectance only 65%, Jaffe, Rank and Wiggins (1955) obtain a resolution of 0·043K at 660K. Since very much higher reflectances are now easily obtainable it is clear that this technique is capable of considerable development.

Multilayers in the ultra-violet and infra-red. There are few materials with high refractive indices which are transparent far into the ultra-violet : although suitable for the visible region, ZnS and CeO_2 begin to absorb sufficiently to make them unsuitable for multilayers at the blue end of the spectrum. Rubidium iodide and cryolite multilayers have been made by Rimbert, Lennuier and Cojan (1955) giving $R = 0{\cdot}70$ and $A = 0{\cdot}05$ at $\lambda = 250\,\text{m}\mu$. Lead chloride and magnesium fluoride yield $R = 0{\cdot}94$ at wavelengths down to 300 mμ (Penselin and Steudel 1955); antimony oxide and cryolite are slightly better, the absorbing cut-off occurring about 8 mμ lower (Barr and Jenkins 1956). Caesium iodide with either cryolite or magnesium fluoride has been used in the region 250–280 mμ the absorption being 2·5–4% for 11 and 13 layers. Caesium bromide and cryolite can be used in the region 230–250 mμ. In these cases it is necessary to protect the soluble alkali halide films by a protective layer of NaF, MgF_2 or paraffin (Steudel and Stotz 1958).

The near infra-red is an especially favourable region since there exist several materials which have refractive indices up to about 6 and which are quite transparent. Germanium and cryolite are satisfactory beyond 1·7 μ until mechanical failure of cryolite sets in at optical thicknesses of the order of 2 μ (i.e. $\lambda/4$ at 8 μ) (Heavens, Ring and Smith 1957). Combinations of tellurium and sodium chloride have been used by Greenler (1955) yielding $R = 0{\cdot}80$ with negligible absorption in the neighbourhood of 12–15 μ.

8.3 *Interference Filters*

8.3.1. *Fabry–Perot type filter.*

The early Fabry–Perot type of interference filter consists of a dielectric spacer layer sandwiched between partly reflecting silver films, deposited on a suitable substrate. The transmission characteristic of such a filter is given approximately by the Airy summation. Maxima occur at integral multiples of wave number $\nu_0 = (2nt)^{-1}$, where n is the refractive index of the spacer and t is the effective thickness (metrical spacer thickness plus path equivalent of phase changes at the dielectric–silver interfaces). The maximum transmittance of the filter is limited by the absorption of the silver layers. The effects of absorption may be reduced by combining the silver layers with suitable multilayer dielectric stacks (Berning and Turner 1957). Considerable improvement in respect of transmittance and bandwidth is obtained by the use of high-reflecting multilayer stacks in place of the silver films. The free spectral range is less than that for the silver + dielectric system since the reflectance of the multilayer stacks decreases away from the maximum. Figure 51 compares the transmission curves for (*a*) the Airy sum, ignoring absorption and (*b*) a five-layer filter for the near infra-red using germanium

(index 4·0) and cryolite (index 1·35). The disadvantage of the limited free spectral range is not serious : the efficiency of the all-dielectric Fabry–Perot filter is so high that unwanted sidebands can easily be removed by auxiliary filtering. The ultimate performance of this type of filter appears to be limited by the inherent inhomogeneities of vacuum-deposited layers. Giacomo, Baumeister and Jenkins (1959) have investigated the effects of errors in the constituent layer thicknesses on the filter performance. The use of a mica spacer (Heavens, Ring and Smith 1957) in place of an evaporated layer leads to a better performance : control of the spacer thickness is difficult although a small range of adjustment of the passband wavelength is obtainable by tilting. Typical figures for maximum transmittance and bandwidth quoted by one manufacturer are : 95% transmittance at 5 mμ bandwidth to 80% transmittance at 1·5 mμ bandwidth (visible spectrum).

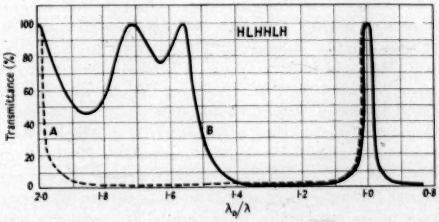


Figure 51. Comparison of Airy curve (A) with that for all-dielectric Fabry–Perot type filter (B). System HLHHLH.

For the near infra-red, filters of this type may be made for wavelengths up to about 10 μ (Greenler 1957). In the region 2–5 μ germanium and cryolite may be used ; bandwidths of the order 1·5% of the passband wavelength are obtained. With tellurium and sodium chloride at $\lambda = 10\ \mu$ a peak transmittance of 40–80% corresponding to bandwidths 0·10 μ to 0·05 μ are possible.

8.3.2. *Double and treble half-wave systems.*

The shape of the transmission band of the Fabry–Perot type of filter is not ideal since one-half of the energy transmitted at one order lies outside the filter half-width. Considerably better performance is obtained with systems of the type

$$\text{stack}/\tfrac{1}{2}\lambda/\text{stack}/\tfrac{1}{2}\lambda/\text{stack} \quad \text{(Smith 1958)}$$

and $\text{stack}/\tfrac{1}{2}\lambda/\text{stack}/\tfrac{1}{2}\lambda/\text{stack}/\tfrac{1}{2}\lambda/\text{stack}$ (Turner *et al.* 1953, Smith 1958).

The double half-wave system can yield a narrow passband such as that of figure 52 : the treble half-wave (described by Turner as a WADI—wide-band all-dielectric interference filter) can give three maxima, as shown in figure 53. At wavelengths around twice the passband wavelength, the double half-wave system may show a fairly steep rise, with continued high transmission to indefinitely long wavelengths, the limit being in practice set by the onset of absorption in the film materials.

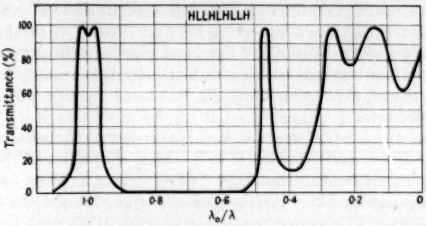


Figure 52. Transmittance characteristic of double half-wave system HLLHLHLLH.

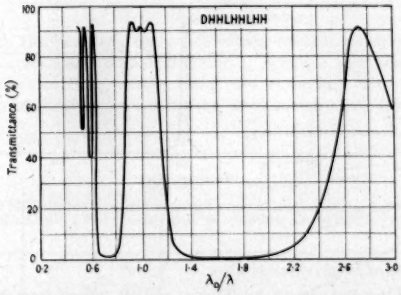


Figure 53. Transmittance characteristic of treble half-wave system DHHLHHLHH.

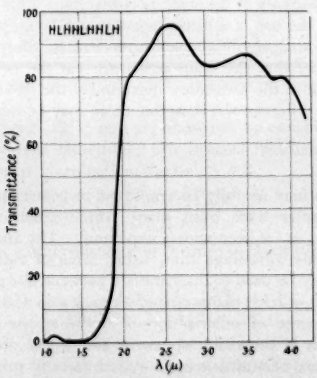


Figure 54. Experimental double half-wave filter for the near infra-red.

Experimental curves of such filters for the infra-red are shown in figure 54. This type of filter is invaluable for suppressing the high-frequency passbands characteristic of any of the narrow-band filters discussed above. The very high absorption of the semiconductor layers used ensures that no visible transmission occurs. If it is desired to cut off at long wavelengths, this can generally easily be effected by the choice of substrate.

A Fabry–Perot type of filter with a variable air-gap as the spacer layer has been constructed by Smith and Heavens (1957) which allows a passband of width of order 0·02 μ to be selected in the range 1·75 to 2·25 μ. When combined with a double half-wave cut-on filter, the transmission characteristic shown in figure 55 is obtained.

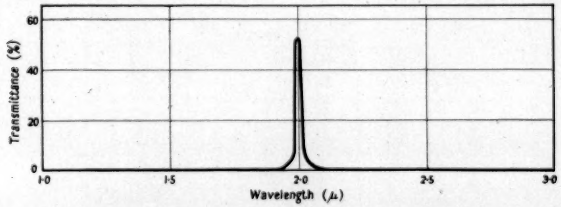


Figure 55. Experimental transmission curve for combination of Fabry–Perot filter with cut-on system. The position of the peak may be varied over $\pm 0{\cdot}25$ μ without loss of shape.

8.3.3. *The frustrated total reflection filter.*

The simple filter introduced by Turner (1950) suffers from the disadvantage of possessing two passbands, arising from the difference in phase changes on reflection of the components of different polarization. This difficulty is overcome by Billings (1950) by the use of a birefringent spacer layer, which brings the passbands together. An analysis of a seven-layer F.T.R. filter with isotropic layers by Kard (1957a) shows that the main passbands for the two components can be brought together. Since the secondary maxima for the two polarized components are found to occur at different wavelengths, these may be eliminated by the use of two such filters.

8.4 *Polarization Devices*

Thin film systems may usefully be employed in polarization devices of various kinds. Several examples have been made which illustrate their potentialities although they have not so far been widely exploited. The analogy with the pile-of-plates polarizer has the advantage that, using films of interference thicknesses, interference effects may be used to discriminate between the polarized components. Thus with two glass ($n = 1{\cdot}53$) plates coated on each side with films of TiO_2, Abelès (1950) produces a degree of polarization of 0·997 at the centre of the visible spectrum ; at the ends of the visible spectrum the degree of polarization is still as high as 0·993. The use of multilayers on a prism face to produce a beam-splitter, described by Banning (1947), enables an average degree of polarization of 0·98 to be

obtained over the visible spectrum for a range of angle of incidence $\pm 5°$ from normal.

The use of thin dielectric films deposited on a totally reflecting prism face, for the production of quarter- and half-wave plates has been examined by Lostis (1959). A single reflection at the hypotenuse of a 45° prism ($n = 1{\cdot}52$) coated with a layer of CeO_2 ($n = 2{\cdot}1$) of thickness $\lambda/3$ produces a relative phase difference of 90°. Over the visible spectrum the variation of the phase difference introduced lies within $\pm 6{\cdot}5°$ of the exact $\lambda/4$ value. For a half-wave layer a film of thickness 0·92 and of index 2·6 is required (TiO_2): in this case the variation over the visible spectrum is $\pm 5°$. Calculations of the dependence of the relative phase difference on film thickness (eqn (27)) show that the performance of such systems is not critically dependent on this quantity.

8.5 *Heat Absorbing Systems*

Systems which absorb nearly the whole of the incident radiation energy are of interest and importance as radiation detectors. A single unsupported metal film absorbs a maximum of 50%, transmitting and reflecting 25% of the incident energy. Systems of three layers—metal/dielectric/metal—have been examined by

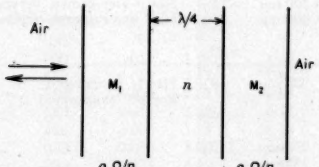


Figure 56.

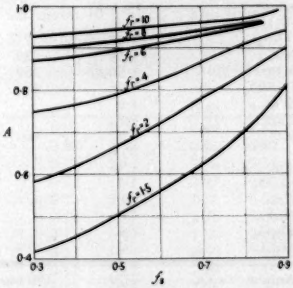


Figure 57.

Hilsum (1954) and Silberg (1957) and are shown to be capable of absorbing up to about 95% of the incident radiation. For normal incidence on the system of figure 56 the absorption is given by

$$A = \frac{4f_s(f_r+1)^2+f_r n^2}{[(f_s+1)(f_r+1)+n^2]^2} \qquad \ldots\ldots(73)$$

where $f_1 = 377/\rho_1$ and ρ_1 is the resistance of the film in ohms per square. Maximum absorption is obtained for

$$n = \frac{(f_r+1)(1-f_s)^{1/2}}{(f_r-1)^{1/2}} \qquad \ldots\ldots(74)$$

for $0 < f_s < 1$ and $f_r > 1$. Figure 57 shows the values of A obtainable for various values of f_r and f_s. For $f_r = 10$ and $f_s = 0{\cdot}8$ the value of n required is approximately 1·5. These values are easily realized in practice.

APPENDIX I

Refractive Indices of Transparent Materials in Thin Film Form

In the table below are collected values obtained for refractive indices of films, formed by thermal evaporation unless otherwise stated. In many cases the actual index depends markedly on the conditions of preparation. Where the effects are known these are mentioned. Values of n are given generally for a wavelength near the centre of the visible spectrum. All the materials listed which are transparent in the visible spectrum show normal dispersion. Where detailed information is available on the dispersion, the appropriate references will be found in §6. An indication is given of the range of wavelength over which the materials are usable in multilayer systems.

Substance	Refractive index	Wavelength (mμ)	Range of transparency from	to	Remarks
CaF_2	1·23–1·26	546	< 150 mμ	15 μ	Poor vacuum
CaF_2	1·36	white			
NaF	1·34	589	< 250 mμ		Bulk index
LiF	1·36–1·37	546	< 110 mμ	10 μ	
Cryolite	1·35–1·39	570	< 200 mμ	10 μ	
AlF_3	1·38–1·39	546			
MgF_2	1·39–1·40	546	230 mμ	10 μ	Aged films
SiO_2	1·44		200 mμ	5 μ	
ThF_4	1·45			> 10 μ	
NaCl	1·54	589	< 200 mμ	> 15 μ	Not water-grown crystals
KBr	1·56	589	600 mμ	30 μ	Not water-grown crystals
$CaSiO_3$	1·68	550			
WO_3	1·8			> 10 μ	
CsBr	1·8		230 mμ	40 μ	
SiO	1·55 to 2·00	vis.	500 mμ	8 μ	Slowly evaporated Rapidly evaporated
Tin oxide	1·96			> 10 μ	
MoO_3	2·0				
CsI	2·0	u. v.	250 mμ	750 μ	
RbI	2·0	u. v.	250 mμ		
Sb_2O_3	2·1	u. v.	290 mμ		
$PbCl_2$	2·0	10 μ	300 mμ	> 14 μ	
	2·2	300 mμ			
Anodic Ta_2O_5	2·20	590 mμ			
ZnS	2·37	546 mμ	400 mμ	14 μ	
Anodic TiO_2	2·40	546 mμ			
CeO_2	2·42	550 mμ			Rapid deposition
Se	2·5		700 mμ	20 μ	
CdS	2·5–2·7	500 mμ	520 mμ	> 14 μ	
Thermal TiO_2	2·73	546 mμ			
Sb_2S_3	3·0	589 mμ	500 mμ	10 μ	
Si	3·5		> 1 μ		
InSb	4·0		3–7 μ		Depends on purity
PbS	4·0	3 μ	3 μ		
Ge	4·3–4·6	2 μ	1·7 μ	20 μ	
Te	4·9	6 μ	3·4 μ	20 μ	
PbSe					
PbTe	5·5		3·8 μ		
Barium stearate	1·49–1·50	white			Deposited from trough by Blodgett technique
Stearic acid	1·54	white			Deposited from trough by Blodgett technique
Haemine	$n = 1·7$, $k = 0·30$	425			Monomolecular layers, 2 mμ thick (Lostis 1955

APPENDIX II

Optical Constants of Opaque Metal Films

Alkali Metals

Measurements made at the quartz–metal interface of films formed by evaporation and subsequently re-melted *in vacuo* between the quartz plates. No details available on pressure in system. The action of re-melting presumably results in a layer with substantially the bulk properties (Ives and Briggs 1936, 1937 a, b).

Wavelength	Sodium		Potassium		Rubidium		Caesium	
(mμ)	*n*	*k*	*n*	*k*	*n*	*k*	*n*	*k*
253·6	0·026	0·62	0·744	0·049	1·031	0·056	0·916	0·143
312·6	0·040	1·02	0·410	0·080	0·814	0·078	0·827	0·174
365·0	0·042	1·44	0·150	0·443	0·496	0·135	0·671	0·233
404·7	0·048	1·56	0·105	0·710	0·275	0·373	0·540	0·320
435·8	0·048	1·80	0·121	0·978	0·181	0·636	0·425	0·438
546·1	0·029	2·32	0·091	1·42	0·157	1·05	0·278	0·950
578·0	0·027	2·59	0·094	1·57	0·164	1·19	0·264	1·123

Thick Annealed Films of Ag, Au, Cu and Al

Measurements by Schulz using the methods given in §5.2.5. The films are annealed until, by the $R_s^2 = R_p$ criterion, they are homogeneous.

Wavelength	Silver		Gold		Copper		Aluminium	
(mμ)	*n*	*k*	*n*	*k*	*n*	*k*	*n*	*k*
400	0·075	1·93	1·45		0·85		0·40	3·92
450	0·055	2·42	1·40	1·88	0·87	2·20	0·49	4·32
500	0·050	2·87	0·84	1·84	0·88	2·42	0·62	4·80
550	0·055	3·32	0·34	2·37	0·72	2·42	0·76	5·32
600	0·060	3·75	0·23	2·97	0·17	3·07	0·97	6·00
650	0·070	4·20	0·19	3·50	0·13	3·65	1·24	6·60
700	0·075	4·62	0·17	3·97	0·12	4·17	1·55	7·00
750	0·080	5·05	0·16	4·42	0·12	4·62	1·80	7·12
800	0·090	5·45	0·16	4·84	0·12	5·07	1·99	7·05
850	0·100	5·85	0·17	5·30	0·12	5·47	2·08	7·15
900	0·105	6·22	0·18	5·72	0·13	5·86	1·96	7·70
950	0·110	6·56	0·19	6·10	0·13	6·22	1·75	8·50

Optical Constants of Ag, Pb and Sn in the Near Infra-red
(Motulevich and Shubin 1957)

λ(μ)	Silver		Lead		Tin	
	n	*k*	*n*	*k*	*n*	*k*
1·03	0·27	7·00	—	—	—	—
1·28	0·36	8·58	1·55	7·0	3·7	8·1
1·71	0·53	11·7	2·1	9·7	3·0	10·7
2·50	0·91	17·2	3·55	14·0	3·6	15·7
3·48	1·65	23·75	5·75	18·5	5·7	20·9
4·38	2·0	29·65	8·0	22·4	6·9	25·2
5·38	2·9	36·9	10·5	26·4	9·0	30·0
6·00	4·3	40·65	12·5	26·8	10·0	32·0
7·09	—	—	12·6	29·5	11·2	34·5

No information on film structure.

References

Abelès, F., 1948 a, *Ann. Phys., Paris*, **3**, 504 ; 1948 b, *C.R. Acad. Sci., Paris*, **226**, 1808 ; 1950, *J. Phys. Radium*, **11**, 310, 403.

Alderson, R. H., and Ashworth, F., 1957, *Brit. J. Appl. Phys.*, **8**, 205.

Avery, D. G., 1950, *Phil. Mag.*, **41**, 1018.

Bailly, R., 1938, *Bull. Cl. Sciences Belg.*, **24**, 791.

Banning, M., 1947, *J. Opt. Soc. Amer.*, **37**, 688, 792.

Bannon, J., 1946, *Nature, Lond.*, **157**, 446.

Barr, W. L., and Jenkins, F. A., 1956, *J. Opt. Soc. Amer.*, **46**, 141.

Baumeister, P. W., and Jenkins, F. A., 1957, *J. Opt. Soc. Amer.*, **47**, 57.

Baumeister, P. W., and Stone, J. M., 1956, *J. Opt. Soc. Amer.*, **46**, 228.

Beattie, J. R., 1955, *Phil. Mag.*, **46**, 235.

Beattie, J. R., and Conn, G. K. T., 1955, *Phil. Mag.*, **46**, 989.

Belk, J. A., Tolansky, S., and Turnbull, D., 1954, *J. Opt. Soc. Amer.*, **44**, 5.

Berning, P. H., 1956, *J. Opt. Soc. Amer.*, **46**, 779.

Berning, P. H., and Turner, A. F., 1957, *J. Opt. Soc. Amer.*, **47**, 230.

Billings, B. H., 1950, *J. Opt. Soc. Amer.*, **40**, 471.

Billings, B. H., and Hyman, M., 1947, *J. Opt. Soc. Amer.*, **37**, 119.

Blaisse, B. S., 1950, *J. Phys. Radium*, **11**, 315.

Blodgett, K. B., and Langmuir, I., 1937, *Phys. Rev.*, **51**, 964.

Bor, J., and Chapman, B. G., 1949, *Nature, Lond.*, **157**, 182.

Bousquet, P., 1956, *Opt. Acta*, **3**, 153 ; 1957, *Ann. Phys., Paris*, **2**, 5.

Bousquet, P., and Delcourt, Y., 1957, *J. Phys. Radium*, **18**, 447.

Bradley, R. S., 1953, *J. Sci. Instrum.*, **30**, 84.

Charlesby, A., and Polling, J. J., 1955, *Proc. Roy. Soc.* A, **227**, 434.

Clegg, P. L., 1952, *Proc. Phys. Soc.* B, **65**, 774.

Clegg, P. L., and Crook, A. W., 1952, *J. Sci. Instrum.*, **29**, 201.

Coogan, C. K., 1957, *Proc. Phys. Soc.* B, **70**, 845.

Cotton, P., 1950, *J. Phys. Radium*, **11**, 375.

Cox, J. T., and Hass, G., 1958, *J. Opt. Soc. Amer.*, **48**, 677.

Cox, J. T., Hass, G., and Rowntree, R. F., 1954, *Vacuum*, **4**, 445.

Crook, A. W., 1948, *J. Opt. Soc. Amer.*, **38**, 954.

David, E., 1939, *Z. Phys.*, **114**, 389.

Dingle, R. B., 1953, *Physica*, **19**, 729.

Ditchburn, R. W., 1952, *Light* (London : Blackie).

Ditchburn, R. W., and Orchard, G. A. J., 1954, *Proc. Phys. Soc.* B, **67**, 608.

Drude, P., 1891, *Wied. Ann.*, **43**, 146.

Dumontet, P., Perrot, M., and Tortosa, J., 1957, *C.R. Acad. Sci., Paris*, **244**, 2488.

Duncan, R. W., and Duncan, R. C., 1913, *Phys. Rev.*, **1**, 294.

Dyson, J., 1958, *Physica*, **24**, 532.

Epstein, L. I., 1952, *J. Opt. Soc. Amer.*, **42**, 806 ; 1955, *Ibid.*, **45**, 360.

Ewald, P., 1916, *Ann. Phys., Lpz.*, **49**, 117.

Fabre, D., and Romand, J., 1956, *C.R. Acad. Sci., Paris*, **242**, 893.

Faust, R. C., 1950, *Phil. Mag.*, **41**, 1238.

Fochs, P. D., 1950, *J. Opt. Soc. Amer.*, **40**, 623.

Försterling, K., 1937, *Ann. Phys., Lpz.*, **30**, 745.

Försterling, K., and Frédericksz, V., 1913, *Ann. Phys., Lpz.*, **40**, 201.

Françon, M., 1952, *Rev. Opt.*, **31**, 65 ; 1953, *Ibid.*, **32**, 349 ; 1954, *Opt. Acta*, **1**, 50.

Frau, C., 1952, *Rev. Opt.*, **31**, 161.

Garnett, J. C. M., 1904, *Phil. Trans. Roy. Soc.*, **203**, 385 ; 1906, *Ibid.*, **205**, 237.

Gates, D. M., Shaw, C. C., and Beaumont, D., 1958, *J. Opt. Soc. Amer.*, **48**, 88.

van Geel, W. C., and Schelen, B. J. J., 1957, *Philips Res. Rep.*, **12**, 240.

Gerharz, R., 1958, *Z. Angew. Phys.*, **9**, 95.

Giacomo, P., 1958, *J. Phys. Radium*, **19**, 307.

Giacomo, P., Baumeister, P. W., and Jenkins, F. A., 1959, *Proc. Phys. Soc.*, **73**, 480.

GIACOMO, P., and JACQUINOT, P., 1952, *J. Phys. Radium*, Suppl. No. 2, 59A.
GILLHAM, E. J., PRESTON, J. S., and WILLIAMS, B. E., 1955, *Phil. Mag.*, **46**, 1051.
GODFREY, G. H., 1957, *Austral. J. Phys.*, **10**, 1.
GRARD, F., 1957 a, *Bull. Cl. Sciences Belg.*, **43**, 641 ; 1957 b, *J. Phys. Radium*, **17**, 414 ; 1957 c, *Bull. Cl. Sciences Belg.*, **43**, 628.
GREENLER, R. G., 1955, *J. Opt. Soc. Amer.*, **45**, 788 ; 1957, *Ibid.*, **47**, 130.
HADLEY, L. N., and DENNISON, D. M., 1947, *J. Opt. Soc. Amer.*, **37**, 451 ; 1948, *Ibid.*, **38**, 483.
HALL, J. F., 1956, *J. Opt. Soc. Amer.*, **46**, 1013 ; 1957, *Ibid.*, **47**, 662.
HALL, J. F., and FERGUSON, W. F., 1955 a, *J. Opt. Soc. Amer.*, **45**, 74 ; 1955 b, *Ibid.*, **45**, 714.
HAMMER, K., 1943, *Z. Tech. Phys.*, **24**, 169.
HARRIS, L., 1955, *J. Opt. Soc. Amer.*, **45**, 27.
HARRIS, L., and LOEB, A. L., 1955, *J. Opt. Soc. Amer.*, **45**, 179.
HARTMAN, R. E., 1954, *J. Opt. Soc. Amer.*, **44**, 192.
HARTMAN, R. E., HARTMAN, R. S., LARSON, K., and BATEMAN, J. B., 1954, *J. Opt. Soc. Amer.*, **44**, 197.
HASS, G., 1949, *J. Opt. Soc. Amer.*, **39**, 532 ; 1952, *Vacuum*, **2**, 331 ; 1955, *J. Opt. Soc. Amer.*, **45**, 945.
HASS, G., HUNTER, W. R., and TOUSEY, R., 1956, *J. Opt. Soc. Amer.*, **46**, 1009 ; 1957, *Ibid.*, **47**, 1070.
HASS, G., RAMSEY, J. B., and THUN, R., 1958, *J. Opt. Soc. Amer.*, **48**, 324.
HASS, G., and SALZBURG, C. D., 1954, *J. Opt. Soc. Amer.*, **44**, 181.
HASS, G., SCHROEDER, H. H., and TURNER, A. F., 1956, *J. Opt. Soc. Amer.*, **46**, 31.
HASS, G., and TURNER, A. F., 1957, *Ergebnisse der Hochvacuumtechnik und der Physik dünner Schichten* (Stuttgart : Wissenschaftliche Verlagsgesellschaft).
HEAVENS, O. S., 1951, *Proc. Phys. Soc.* B, **64**, 419 ; 1955, *Optical Properties of Thin Solid Films* (London : Butterworths Scientific Publications) ; 1959, *Lab. Pract.*, in the press.
HEAVENS, O. S., and KELLY, J. C., 1958, *Proc. Phys. Soc.*, **72**, 906 ; 1959, *Opt. Acta*, **6**, 339.
HEAVENS, O. S., RING, J., and SMITH, S. D., 1957, *Spectrochim. Acta*, **10**, 179.
HEAVENS, O. S., and SMITH, S. D., 1957, *J. Opt. Soc. Amer.*, **47**, 469.
VAN HEEL, A. C. S., and WALTHER, A., 1958, *Opt. Acta*, **5**, 47.
HERMANSEN, A., 1955, *K. Danske Vidensk. Selsk., Mat.-Fys. Medd.*, **30**, No. 6.
HERPIN, A., 1947, *C.R. Acad. Sci., Paris*, **225**, 182.
HILSUM, C., 1954, *J. Opt. Soc. Amer.*, **44**, 188.
HODGSON, J. N., 1955, *Proc. Phys. Soc.* B, **68**, 593.
HOLLAND, L., and SIDDALL, G., 1953, *Vacuum*, **3**, 375.
HULDT, L., and STAFLIN, T., 1959, *Opt. Acta*, **6**, 27.
HUTCHINSON, A., 1907, *Min. Mag.*, **14**, 199.
IVES, H. E., and BRIGGS, H. B., 1936, *J. Opt. Soc. Amer.*, **26**, 238 ; 1937 a, *Ibid.*, **27**, 395 ; 1937 b, *Ibid.*, **27**, 181.
JAFFE, J. H., RANK, D. H., and WIGGINS, T. A., 1955, *J. Opt. Soc. Amer.*, **45**, 636.
JARRETT, A. H., 1952, *Nature, Lond.*, **169**, 790.
JENESS, J. R., 1956, *J. Opt. Soc. Amer.*, **46**, 157.
JOHNSON, W. C., 1955, *Organic Reagents for Metals* (London : Hopkins and Williams).
KANDARE, S., and FABRE, D., 1956, *C.R. Acad. Sci., Paris*, **242**, 1150.
KARD, P. G., 1956, *Soviet Physics, Doklady*, **1**, 256 ; 1957 a, *Izvest. Akad. Nauk Eston., S.S.S.R.*, **6**, 344 ; 1957 b, *Opt. i Spekt.*, **2**, 236, 245.
KELLY, J. C., 1958 a, *Opt. Acta*, **5**, 75 ; 1958 b, *Ph.D. Thesis*, University of Reading.
KOEHLER, W. F., 1953, *J. Opt. Soc. Amer.*, **43**, 738 ; 1955, *Ibid.*, **45**, 943.
KRAUS, H. L., 1949, *Elect. Engng*, **68**, 767.
KRAUTKRÄMER, J., 1938, *Ann. Phys., Lpz.*, **32**, 537.
KRONIG, R. DE L., 1931, *Proc. Roy. Soc.* A, **133**, 255.
KRYLOVA, T. N., 1958, *Opt. i Spekt.*, **4**, 217.
KUNZE, C., 1958, *Ann. Phys., Lpz.*, [7], **1**, 165.
KUWABARA, G., and ISIGURO, K., 1952, *J. Phys. Soc. Japan*, **7**, 72.
LEBERKNIGHT, C. E., and LUSTMAN, B., 1939, *J. Opt. Soc. Amer.*, **29**, 59.

LISITSA, M. P., and TSVELYKH, N. G., 1957, *Opt. i Spekt.*, **2**, 674.
LISSBERGER, P. H., and RING, J., 1955, *Opt. Acta*, **2**, 42.
LOCKHART, L. B., and KING, P., 1947, *J. Opt. Soc. Amer.*, **37**, 689.
LOSTIS, P., 1955, *C.R. Acad. Sci., Paris*, **240**, 2130 ; 1959, *Rev. Opt.*, **38**, 1.
MCLAURIN, R. C., 1905, *Proc. Roy. Soc.* A, **76**, 149.
MALÉ, D., 1950, *J. Phys. Radium*, **11**, 332 ; 1952, *Thèse*, Université de Paris.
MALÉ, D., and ROUARD, P., 1953, *J. Phys. Radium*, **14**, 584.
MALÉ, D., and TROMPETTE, J., 1957, *J. Phys. Radium*, **18**, 128.
MATTUCK, R. D., 1956, *J. Opt. Soc. Amer.*, **46**, 615, 621.
MATTUCK, R. D., PETTI, R. D., and BATEMAN, J. B., 1956, *J. Opt. Soc. Amer.*, **46**, 782.
MAYER, H., 1950, *Physik dünner Schichte* (Stuttgart : Wissenschaftliche Verlagsgesellschaft).
MORITA, N., 1956, *J. Phys. Soc. Japan*, **11**, 957 ; 1957, *Ibid.*, **12**, 1142.
MOTULEVICH, G. P., and SHUBIN, A. A., 1957, *Opt. i Spekt.*, **2**, 633.
MUCHMORE, R., 1948, *J. Opt. Soc. Amer.*, **38**, 20.
MURMANN, H., 1933, *Z. Phys.*, **80**, 161 ; 1936, *Ibid.*, **101**, 643.
MURRAY, A. E., 1956, *J. Opt. Soc. Amer.*, **46**, 790.
O'BRYAN, H. M., 1936, *J. Opt. Soc. Amer.*, **26**, 122.
PENSELIN, S., and STEUDEL, A., 1955, *Z. Phys.*, **142**, 21.
PHILIP, R., 1956, *C.R. Acad. Sci., Paris*, **243**, 365.
PHILIP, R., and TROMPETTE, J., 1957, *J. Phys. Radium*, **18**, 92.
RANK, D. H., and BENNETT, H. E., 1955, *J. Opt. Soc. Amer.*, **45**, 69.
REIMER, L., 1957, *Optik*, **14**, 83.
REUTER, G. E. H., and SONDHEIMER, E. H., 1948, *Proc. Roy. Soc.* A, **195**, 336.
RIMBERT, G., LENNUIER, R., and COJAN, J.-L., 1955, *C.R. Acad. Sci., Paris*, **241**, 1560.
RING, J., BEER, R., and HEWISON, V., 1958, *J. Phys. Radium*, **19**, 321.
ROOD, J. L., 1949, *J. Opt. Soc. Amer.*, **39**, 854.
ROTHEN, A., 1945, *Rev. Sci. Instrum.*, **16**, 26.
ROTHEN, A., and HANSON, M., 1949, *Rev. Sci. Instrum.*, **20**, 66.
ROUARD, P., 1937, *Ann. Phys., Paris*, **7**, 291.
ROUARD, P., MALÉ, D., and TROMPETTE, J., 1953, *J. Phys. Radium*, **14**, 587.
SANDELL, E. B., 1944, *Colorimetric Determination of Traces of Metals* (New York : Interscience).
SANFORD, B. P., 1958, *J. Opt. Soc. Amer.*, **48**, 482.
SCHOPPER, H., 1951, *Z. Phys.*, **130**, 565 ; 1952 a, *Ibid.*, **132**, 146 ; 1952 b, *Ibid.*, **131**, 215.
SCHULZ, L. G., 1949, *J. Chem. Phys.*, **17**, 1153 ; 1950, *J. Opt. Soc. Amer.*, **40**, 690 ; 1951, *Ibid.*, **41**, 261 ; 1954 a, *Ibid.*, **44**, 357 ; 1954 b, *Ibid.*, **44**, 540 ; 1957, *Advanc. Phys.* (*Phil. Mag.* Suppl.), **6**, 102.
SCHULZ, L. G., and SCHEIBNER, E. J., 1950, *J. Opt. Soc. Amer.*, **40**, 761.
SCHULZ, L. G., and TANGHERLINI, F. R., 1954, *J. Opt. Soc. Amer.*, **44**, 362.
SCOTT, G. D., MCLAUCHLAN, T. A., and SENNETT, R. S., 1950, *J. Appl. Phys.*, **21**, 843.
SENNETT, R. S., and SCOTT, G. D., 1950, *J. Opt. Soc. Amer.*, **40**, 203.
SHALIMOVA, K. V., 1951, *Dokl. Akad. Nauk, S.S.S.R.*, **80**, 587.
SILBERG, P. A., 1957, *J. Opt. Soc. Amer.*, **47**, 575.
SINELNIKOV, K. D., SHKLIAREVSKY, I. N., and VLASENKO, N. A., 1957, *Opt. i Spekt.*, **2**, 651.
SIVUKHIN, D. V., 1956, *Soviet Physics, JETP*, **3**, 269.
SMITH, S. D., 1958, *J. Opt. Soc. Amer.*, **48**, 43.
SMITH, S. D., and HEAVENS, O. S., 1957, *J. Sci. Instrum.*, **34**, 492.
SMITH, S. D., and MOSS, T. S., 1958, *J. Sci. Instrum.*, **35**, 105.
STEUDEL, A., 1957, *Naturwissenschaften*, **44**, 249.
STEUDEL, A., and STOTZ, S., 1958, *Z. Phys.*, **151**, 233.
STRACHEN, C., 1933, *Proc. Camb. Phil. Soc.*, **29**, 116.
SUKHANOVSKY, V. V., 1956, *Dokl. Akad. Nauk, S.S.S.R.*, **106**, 226.
THELEN, A., 1956, *Optik*, **13**, 537.
TOLANSKY, S., 1948, *Multiple Beam Interferometry* (Oxford : University Press) ; 1951, *J. Electrodep. Tech. Soc.*, **27**, 1.
TRAUB, A. C., 1956, *J. Opt. Soc. Amer.*, **46**, 999.
TRAUB, A. C., and OSTERBERG, H., 1957, *J. Opt. Soc. Amer.*, **47**, 62.

TROMPETTE, J., 1959, *C.R. Acad. Sci., Paris*, **248**, 207.
TURNER, A. F., 1950, *J. Phys. Radium*, **11**, 444.
TURNER, A. F., and HOPKINSON, H. R., 1953, *J. Opt. Soc. Amer.*, **43**, 819.
TURNER, A. F., *et al.*, 1953, *Infra-red Transmission Filters*, Bausch and Lomb Technical Reports, Nos. 1–6.
VAŠIČEK, A., 1940, *Phys. Rev.*, **57**, 925; 1947, *J. Opt. Soc. Amer.*, **37**, 145; 1950, *J. Phys. Radium*, **11**, 342; 1951, *Act. Acad. Sci. Nat., Moravo-Silesiacae*, **23**, 355.
VASILKOVSKY, A. A., 1957, *Opt. i Spekt.*, **2**, 229.
WEAVER, C., and BENJAMIN, P., 1956, *Nature, Lond.*, **177**, 1030.
WEINSTEIN, W., 1947, *J. Opt. Soc. Amer.*, **37**, 576; 1954, *Vacuum*, **4**, 3.
WEISS, K., 1948, *Z. Naturf.*, **3a**, 143.
WEISSKIRCHNER, W., 1951, *Z. Naturf.*, **6a**, 509.
WINTERBOTTOM, A. B., 1946, *Trans. Farad. Soc.*, **42**, 487.
WOLTER, H., 1937, *Z. Phys.*, **105**, 269.
YOUNG, L., 1958, *Proc. Roy. Soc.* A, **244**, 41.
YOUNG, N. O., and KOWAL, J., 1959, *Nature, Lond.*, **183**, 104.

GROUP THEORY IN SOLID STATE PHYSICS

By D. F. JOHNSTON

Theoretical Physics Division, Atomic Energy Research Establishment, Harwell, Didcot, Berks.

CONTENTS OF PART 1

	PAGE
Introduction	67
§ 1. Vector spaces—function spaces, product spaces, sub-spaces	68
§ 2. Scalar products in vector spaces—the axioms of unitary geometry	71
§ 3. Operators on vectors—linear, adjoint, Hermitian, unitary	73
§ 4. Sets of operators—groups and group algebras, reducibility of spaces, Schur's lemma	80
§ 5. Symmetry in quantum mechanics—the postulate of irreducibility and its consequences	84
§ 6. Matrix representations of groups—characters, unitary representations	88
§ 7. The group-theoretical simplification of eigenvalue calculations	101

CONTENTS OF PART 2

Introduction	106
§ 1. Lattices and space groups	107
§ 2. The symmetry group of the Dirac Hamiltonian	110
§ 3. Wigner time-reversal symmetry	121
§ 4. Group theory and the many-electron problem	128
§ 5. Irreducible representations of space groups	138
Acknowledgments	147
Appendix I	148
Appendix II	150
Appendix III	152
References	152

Abstract.

Part 1. The concept of a vector space irreducible under a set of operators is developed from first principles, and then introduced into the many-particle formalism of quantum mechanics by means of an explicit postulate of irreducibility.

The calculus of the irreducible matrix representations of finite groups is developed *ab initio*, including the theory of characters and projection operators. The use of this calculus to simplify eigenvalue calculations is explained in detail.

Part 2. The structure of the general crystal space group, including glide-planes and screw-axes, is discussed briefly. The theory is developed of the symmetry group of a many-electron system with spin–orbit coupling, using the Dirac formalism. A detailed discussion is given of Wigner's time-reversal theorems for a many-electron system, including the character tests for time-reversal degeneracy. A general theory of the permutation symmetry of a many-electron system is developed, and shown to contain the Dirac vector model as a special case.

A new treatment is given of the theory of the irreducible representations of space groups, including the double space groups and Herring's time-reversal theorems.

PART 1

Introduction

In writing this article, two kinds of reader have been borne in mind: on the one hand, the new post-graduate student approaching the subject for the first time; and on the other, the reader, already familiar with group-theoretical ideas, seeking a convenient exposition of the applications to solid-state physics. It follows that both kinds of reader will find sections of the article which can be omitted, at least on first reading. In particular, the new student is advised not to concern himself at first with the proofs of theorems, if he can understand the contents of the theorems.

Group theory differs from physical theories like, say, electromagnetism or thermodynamics, in that the basic concepts of physical theories are immediately accessible to the imagination, either through pointer readings on instruments, or directly as sense perceptions. The basic concepts of group theory, on the other hand, are essentially abstractions from the external world. It follows that group-theoretical ideas must be expressed in terms of a very well defined vocabulary if we are to know exactly what we are doing and what every statement we make really means. We cannot appeal to instruments or to our senses for reassurance that we are not talking nonsense. (The author himself is a simple-minded physicist who has found himself talking nonsense on too many occasions.) This is the reason for an apparent pre-occupation with definitions, and necessary and sufficient conditions, in the first part of the article.

The basic ideas of group theory are really very simple. On the one hand we consider a set of vectors denoted by $\{X\}$, and on the other, a set of operators denoted by $\{A\}$, such that for every operator A in $\{A\}$ and vector X in $\{X\}$, the vector AX, arising from the operation A on X, can be expressed as a linear combination of vectors in $\{X\}$. We have a simple example if we take $\{X\}$ to be a set of basis vectors in our familiar three-dimensional Cartesian space, and $\{A\}$ the set of all rigid rotations about the origin.

A set of vectors $\{X\}$ becomes a vector space over a field of numbers if we require $\{X\}$ to contain all linear combinations over these numbers, of any collection of vectors in $\{X\}$. We can then speak of $\{X\}$ as being invariant under $\{A\}$, if A in $\{A\}$, X in $\{X\}$, implies (AX) in $\{X\}$. This is clearly true of our example above, where the field is the field of real numbers.

We shall be particularly concerned with situations in which we consider a vector space $\{X\}$ and set of operators $\{A\}$, so related, that given any non-zero vector Y in $\{X\}$, the vector space generated from the set $\{AY\}$, A in $\{A\}$, is identical to the space $\{X\}$. In this case we say that the space $\{X\}$ is irreducible with respect to $\{A\}$.

This collection of primitive ideas is complete if the set of operators $\{A\}$ is a group, that is:

(1) $\{A\}$ contains the unit operator I.

(2) A in $\{A\}$ implies A^{-1} in $\{A\}$, where A^{-1} is the reciprocal of A.

(3) $\left.\begin{matrix} A \text{ in } \{A\} \\ B \text{ in } \{A\} \end{matrix}\right\}$ implies $\begin{matrix} AB \text{ in } \{A\}, \\ BA \text{ in } \{A\}. \end{matrix}$

This group property and the concepts of invariance and irreducibility are the foundation stones of the whole edifice of the theory of group representations.

The article has been arranged in two parts: in the first part we develop the concepts, relations and principles of the subject; in the second part we consider the applications to solid-state problems. It is suggested that these two parts be read in parallel. A bibliography will be found in the Appendix.

§1. Vector Spaces

As we indicated in the Introduction, the concept of a vector space is basic for our purpose. In this first section we abstract from our familiar everyday world the general concept of a finite dimensioned space. We then give examples of the kind of spaces we shall require.

The first step in developing our ideas is to generalize the concept of a vector space, starting from the familiar language with which we describe the position of objects around us. This language is based on the concept of a three-dimensional vector space in which the coordinates of the vectors are real numbers. We are familiar with the associated vector calculus in which we consider linear combinations of vectors, for example

$$X = aY + bZ,$$

where X, Y, Z are vectors and a and b real numbers. Let us describe these familiar ideas in a language which will enable us to generalize the concept of a vector space. We say then that our vector space is a set S whose infinitely many members are called vectors. This set S and the associated vector calculus are such that:

(1) Given vectors X and Y in S we have a prescription for forming a new vector in S, called the sum of X and Y and denoted by $(X+Y)$ or $(Y+X)$, these two expressions being regarded as identical. The prescription is of course that the coordinates of the sum vector $(X+Y)$ are the sums of the respective coordinates of the vectors X and Y.

(2) Given any vector X and a number c (for our purpose we generalize to complex numbers), we have a prescription for forming a new vector in S denoted by cX, the prescription being that the coordinates of the vector cX are the coordinates of X each multiplied by the complex number c. Combining (1) and (2), given vectors X and Y and complex numbers a and b, we can form the vector $(aX+bY)$ in S; for we have a prescription for calculating the coordinates of the vector $(aX+bY)$ from those of X and Y.

Now consider three non-coplanar vectors X, Y and Z in our familiar three-dimensional space. It is evident that the only solution in the numbers a, b, c to the equation

$$aX + bY + cZ = 0,$$

is $a = b = c = 0$. Instead of speaking of vectors as being non-coplanar, we shall speak of vectors as being linearly independent. We shall say that the vectors $X_1, X_2, \ldots, X_n$ are linearly independent if the relation

$$a_1 X_1 + a_2 X_2 + \ldots + a_n X_n = 0$$

implies $a_1 = a_2 = \ldots = a_n = 0$. The dimensionality of a vector space S is the maximum number of linearly independent vectors that can be found in S. In

principle a vector space is defined by a given set of vectors $\{X\}$, not necessarily linearly independent; for a vector space S can be generated from $\{X\}$ by adding to the set $\{X\}$ all linear combinations over the complex numbers of the given vectors in $\{X\}$. In this sense we speak of a set of vectors $\{X\}$ spanning the vector space S. A set of basis vectors $\{e_r\}$ in S is a set of linearly independent vectors e_r in S with the property that every vector X in S can be expressed as a linear combination of the basis vectors; that is, $X = \Sigma_r e_r X_r$ where the numbers X_r are the coordinates of X referred to the basis $\{e_r\}$. On the other hand, the concept of a vector X contained in a vector space S does not require reference to an explicit basis.

This description of our familiar three-dimensional position vector space has been deliberately made so general that the restriction to three dimensions and real coordinates does not appear. In fact, our familiar three-dimensional space is a particular example of the concept of a vector space over the field of complex numbers; that is, a vector space in which the coordinates X_r of a vector X referred to some basis $\{e_r\}$ are in general complex numbers. A very elegant treatment of angular momentum in classical mechanics (see Corben and Stehle 1950) can be given by introducing a complex basis into our familiar position space. This is because rotations take a specially simple form in such a basis: to demonstrate this, let e_x, e_y, e_z denote the usual orthogonal basis, then a rotation $R(\theta)$ about the z axis e_z has the property

$$Re_x = e_x \cos\theta + e_y \sin\theta, \quad Re_y = -e_x \sin\theta + e_y \cos\theta, \quad Re_z = e_z.$$

If, instead, we refer the rotation $R(\theta)$ to the basis

$$e_1 = \frac{1}{\sqrt{2}}(e_x + ie_y), \quad e_2 = \frac{1}{\sqrt{2}}(e_x - ie_y), \quad e_3 = e_z,$$

we find the simple relations

$$Re_1 = e_1 \exp(-i\theta), \quad Re_2 = e_2 \exp i\theta, \quad Re_3 = e_3.$$

Function spaces.

One kind of vector space with which we shall be concerned is a function space; that is, a vector space whose basis vectors form a set of linearly independent functions of some argument. The general vector in such a space is an arbitrary linear combination of the basis functions over the complex numbers. For example, the three functions $xy/r^2, yz/r^2, zx/r^2$ span a three-dimensional function space. We shall regard as vectors the wave functions of quantum mechanics.

Product spaces.

Given two vector spaces, say S_m of m-dimensional basis $\{u_p\}$ $(p = 1 \ldots m)$ and S_n of n-dimensional basis $\{v_r\}$ $(r = 1 \ldots n)$, we can regard the $n \times m$ formal products $u_p v_r$ $(p = 1 \ldots m, r = 1 \ldots n)$ as the basis vectors of an $m \times n$-dimensional space which we call the product of S_m and S_n and denote by $S_m \times S_n$. The general vector X in $S_m \times S_n$ is then

$$X = \sum_{p,r} (u_p v_r) X_{pr}.$$

It is evident that this constructional procedure can be extended to the product of any number of spaces. For example, consider the product of three two-dimensional spaces S(1), S(2), S(3). Let the basis vectors of S(1) be $\alpha(1)$ and $\beta(1)$, of S(2), $\alpha(2)$ and $\beta(2)$, and of S(3), $\alpha(3)$ and $\beta(3)$. The basis vectors of the eight-dimensional product space $S(1) \times S(2) \times S(3)$ are then the formal products

$$\begin{matrix} \alpha(1)\alpha(2)\alpha(3), & \alpha(1)\alpha(2)\beta(3), & \beta(1)\beta(2)\alpha(3), \\ \beta(1)\beta(2)\beta(3), & \alpha(1)\beta(2)\alpha(3), & \beta(1)\alpha(2)\beta(3), \\ & \beta(1)\alpha(2)\alpha(3), & \alpha(1)\beta(2)\beta(3). \end{matrix}$$

We shall use this example of a vector space again. As a second example of a product space we can combine a number of function spaces, the basis vectors being products of functions.

Let us pause here for a moment, and note that this concept of a product space is only accessible to our imagination, via an abstract definition of a vector space, derived from the familiar world. Product spaces are of great importance in group theory and quantum mechanics.

In order to describe a particle with spin in quantum mechanics we combine a function space with a spin-space. Let $\{e_n\}$ denote a basis in a spin-space, and let $\{\phi_r\}$ denote a set of basis functions in a function space, then the general vector Ψ in the corresponding product space is

$$\Psi = \sum_{n,r} e_n \phi_r c_{nr} = \sum_n e_n \sum_r \phi_r c_{nr} = \sum_n e_n \psi_n,$$

where the functions ψ_n are arbitrary functions in the function space. We shall call such a product space a one-particle space. In a many-particle problem in quantum mechanics we assign a one-particle space to each particle, and then form a many-particle product space by formally multiplying these one-particle spaces together.

Change of basis.

Given a vector space S defined by a set $\{u_r\}$ of linearly independent vectors, it does not follow that these vectors form the most convenient basis vectors in S. Let $\{v_j\}$ denote some other basis set in S, then there exists a linear transformation between $\{u_r\}$ and $\{v_j\}$; that is, there exist square arrays of complex numbers P_{rj}, Q_{ks} such that $v_j = \Sigma_r u_r P_{rj}$, $u_s = \Sigma_k v_k Q_{ks}$. We are familiar with this idea in our three-dimensional position space.

Sub-spaces.

The concept of a set of vectors defining a vector space leads to the idea of a sub-space of a given space. Let S denote a given space, and let $\{X\}$ denote a set of vectors, such that each member X of the set $\{X\}$ is a vector in S. The space spanned by $\{X\}$ is a sub-space of S, say S′. In general, not every vector in S is a member of S′, unless we have chosen S′ to coincide with S. The space S′ could be the zero space, that is, the empty set. We say that S′ is a proper sub-space of S if $S' \neq S$ and $S' \neq 0$. A simple example of a proper sub-space is that of a plane containing the origin in our three-dimensional position space.

Complementary sub-space.

Given a space S and a sub-space S′ of S, a second sub-space S″ of S is complementary to the sub-space S′, if S″ is such that S′ and S″ contain no non-zero vector in common but every vector Z in S can be expressed as a sum of a vector X in S′ and a vector Y in S″ ; that is, $Z = X + Y$. We can demonstrate that the complementary sub-space S″ so defined may not be unique, for in our example of a plane through the origin as a sub-space in our three-dimensional space, any vector through the origin, not contained in the plane, forms the basis of a one-dimensional sub-space complementary to the plane. We shall return to this point in § 2.

§ 2. Scalar Products in Vector Spaces

The formalism of group theory and quantum mechanics makes use of a generalization of the concept of a scalar product between a pair of vectors, an idea with which we are already familiar in our three-dimensional language. We choose our definitions and abstractions in a form which will be fruitful for our purpose. If our physical universe had been of a different kind, a different choice might have been relevant.

In our three-dimensional position vector space, to every pair of vectors X and Y there corresponds a number $(X.Y)$, the prescription for calculating the scalar product $(X.Y)$ being the familiar formula $(X.Y) = |X||Y|\cos\theta$. These scalar products satisfy the linear relations

$$(X.Y) = (Y.X), \quad (aY+bZ).X = a(Y.X)+b(Z.X),$$

for arbitrary vectors X, Y, Z and real numbers a and b. Now having abstracted these relations from our familiar world, we generalize them, and then use the new relations to define a generalized scalar product. We say then, that a vector space S is a unitary space, if to every *ordered* pair of vectors X and Y in S, there corresponds a complex number $(X.Y)$, the scalar product of X and Y in that order, these products satisfying the axioms

$$(X.Y) = (Y.X)^*, \; (X.aY) = a(X.Y),$$
$$(aY+bZ).X = a^*(Y.X)+b^*(Z.X),$$

for arbitrary vectors X, Y, Z in S, and complex numbers a and b. (It is the statistical nature of quantum mechanical prediction that imposes upon us this choice of a unitary geometry.) Our familiar three-dimensional geometry of position is a unitary geometry over the real numbers.

It is sufficient to know the scalar products $(u_r.u_s) = (u_s.u_r)^*$ for every pair of vectors in a given basis $\{u_r\}$ of a space S ; for given $X = \Sigma_r u_r X_r$ and $Y = \Sigma_s u_s Y_s$ in S, it follows from the axioms that

$$(X.Y) = \sum_{r,s} (u_r.u_s) X_r^* Y_s.$$

We can use our formal definition of a scalar product to define scalar products between vectors in *different* spaces, provided these spaces are of the same dimensions. If $X = \Sigma_r u_r X_r$ is a vector in S′ and $Y = \Sigma_s v_s Y_s$ is a vector in S″, then $(X.Y)$ is defined, if we define what we mean by the symbols $(u_r.v_s)$. This is the essence of Dirac's bra and ket notation.

Orthogonality and normalization.

Once we have defined scalar products, we can introduce the general concept of orthogonality. Vectors X and Y in a unitary space are said to be orthogonal if $(X.Y) = 0$. A vector X is said to be normalized to unity if $(X.X) = 1$. In practice it may be convenient to normalize to some positive number other than unity. Given that $X = aY$, where the complex number a is to be chosen so that X is normalized, we have the condition $|a|^2 = (Y.Y)^{-1}$; that is, the phase of the complex number a is unconstrained.

One of the most fruitful concepts derived from these ideas is, of course, that of an orthonormal basis in a vector space. The basis $\{u_r\}$ of a space S is orthonormal if $(u_r.u_s) = \delta_{rs}$. We have then the simple formulae

$$(X.Y) = \Sigma_r X_r^* Y_r, \quad X = \Sigma_r u_r(u_r.X), \quad \text{since} \quad X_r = (u_r.X).$$

Scalar products in function spaces.

It is not necessary to work in terms of a basis if we have a prescription for calculating the complex numbers $(X.Y)$ directly from the given vectors X and Y. If the functions ϕ of a function space are such that the integral square $\int|\phi|^2 d\tau$ over some domain of the argument is finite for all vectors in the function space, we can define the scalar product between two vectors (functions) ψ and ϕ by the prescription

$$(\psi.\phi) = \int \psi^* \phi d\tau,$$

a formula well known in Schrödinger's wave mechanics. It is evident that this prescription satisfies our axioms.

Scalar products in product spaces.

The scalar product between vectors

$$X = \sum_{p,r} (u_p v_r) X_{pr} \quad \text{and} \quad Y = \sum_{q,s} (u_q v_s) Y_{qs}$$

in our product space $S_m \times S_n$ is defined to be

$$(X.Y) = \sum_{p,q,r,s} (u_p.u_q)(v_r.v_s) X_{pr}^* Y_{qs}.$$

Scalar products between vectors $\Psi = \Sigma_n e_n \psi_n$ and $\Phi = \Sigma_m e_m \phi_m$ describing particles with spin are evaluated according to the prescription

$$\Psi.\Phi = \sum_{n,m} (e_n.e_m) \int \psi_n^* \phi_m d\tau.$$

If the basis $\{e_n\}$ of the spin space is chosen to be orthonormal, this simplifies to

$$\Psi.\Phi = \sum_n \int \psi_n^* \phi_n d\tau.$$

Orthogonal complementary sub-space.

We can now introduce the concept of orthogonality into our discussion of sub-spaces and complementary sub-spaces. We noted that in general the sub-space S″ complementary to a given sub-space S′ of a vector space is not unique. We introduce the concept of an orthogonal complementary sub-space, by imposing

the further condition on S″ that X in S′, Y in S″ implies $X.Y = 0$. In our example a vector through the origin normal to the given plane forms the basis of an orthogonal complementary sub-space relative to the plane.

§ 3. Operators

We come now to the most fascinating concepts—the operators on vector spaces. Whereas vector spaces are relatively docile, the operators are adventurers whose actions can lead to unexpected results, unless we are very careful in our definitions. We are concerned here with operators, say A, which have the power to change a given vector X into some other vector Y. We write then $Y = AX$. Let us consider some examples of such operators.

Position operators.

Let $\{a_r\}$ denote a set of basis vectors in our three-dimensional position space. Let α denote a rotation under which the vector a_r is carried into the vector $a_r' = \alpha a_r$. The general position vector $X = \Sigma_r a_r X_r$ is carried into the position vector $X' = \alpha X$, where by definition $\alpha X = \Sigma_r (\alpha a_r) X_r$. Note that this definition implies $\alpha(aX) = a(\alpha X)$, where a is a number. Given a position operator, we can define a corresponding substitutional operator on functions.

Substitutional operators.

Let $\psi(X)$ be a function of a position vector X, and let α denote a position operator with a reciprocal α^{-1}; that is, if $X' = \alpha X$, then $\alpha^{-1} X' = X$, for arbitrary vector X. We can now define a new function $\phi(X)$ by the relations

$$\left.\begin{aligned} \phi(X') &= \psi(X) \\ X' &= \alpha X \end{aligned}\right\} \text{ true for all } X.$$

Since α^{-1} exists, we can write this definition in the form

$$\phi(X) = \psi(\alpha^{-1} X),$$

and regard $\phi(X)$ as the result of operating on $\psi(X)$ with an operator, which we denote by the same symbol α. That is, we write $\phi(X) = \alpha\psi(X)$. The reasons for choosing this particular form of definition will appear later.

Permutation operators.

The permutation operators required in the group theory of a many-particle problem in quantum mechanics are operators on product spaces. Consider a product space S constructed by multiplying formally together N vector spaces S_r ($r = 1 \ldots N$), each of the same kind, in particular each of the same dimensionality. We can take as our basis in S a set of basis vectors, each of which is a formal product (Hartree product) of N vectors, one from each space S_r. These basis vectors can be obtained one from another, either by permuting the labels of the component spaces S_r, or by permuting the elementary vectors in the Hartree product among the spaces S_r. It is evident that there is a one-to-one correspondence between the two definitions of such an operator. For example, let $X(1)\,Y(2)\,Z(3)$ denote a Hartree product belonging to the product of three similar spaces, then we can permute the vectors X, Y, Z among the spaces, or we can

relabel the spaces, by permuting the numbers 1, 2, 3. We now choose our definitions of these operators in such a way that we can use the same symbol for the operator in either case. We can do this if we allow the position of an elementary vector in a Hartree product to define the component space to which it belongs, then $X(1)\,Y(2)\,Z(3)$ is written XYZ. A permutation operator is represented by, for example, the symbol

1	2	3
2	3	1

with the alternative, but equivalent, interpretations :

(a) the elementary vector in S_1 is carried into S_2, that in S_2 is carried into S_3, and the vector in S_3 is carried into S_1 ;

(b) the space S_1 is relabelled S_2, the space S_2 is relabelled S_3, while the space S_3 is relabelled S_1.

Strictly speaking, in quantum mechanics we should relabel the spaces, but, as we have seen, the two procedures are exactly equivalent in this context.

The reader may find it helpful to experiment with these permutation operators on the basis vectors of the eight-dimensional product space, used as an example in the section on product spaces.

Linear operators.

With certain exceptions, all the operators with which we shall be concerned possess the linear property, namely

$$A(aX+bY) = a(AX)+b(AY)$$

for arbitrary vectors X and Y and complex numbers a and b. We shall see that this property distinguishes between the devils and the angels in our operator society. The linear property cannot be assumed without discussion ; for example, in our three-dimensional space consider a translation operator defined by the relation $TX = X+t$ where X is an arbitrary vector in the space and t is a constant vector, then

$$\begin{aligned} T(aX+bY) = TZ, \qquad & Z = aX+bY, \\ = Z+t &= aX+bY+t ; \end{aligned}$$

that is, the operator T so defined is *not* a linear operator, for if T were linear we should have

$$T(aX+bY) = a(TX)+b(TY) = aX+bY+(a+b)\,t.$$

On the other hand, if we define the translation operator T as a substitutional operator in a function space by the relation

$$T\psi(\mathbf{r}) = \psi(\mathbf{r}-t) \text{ for all } r,$$

then

$$\begin{aligned} T\{a\psi(\mathbf{r})+b\phi(\mathbf{r})\} = T\theta(\mathbf{r}) &= a\psi(\mathbf{r}-t)+b\phi(\mathbf{r}-t) \\ &= aT\psi(\mathbf{r})+bT\phi(\mathbf{r}) ; \end{aligned}$$

that is, T so defined is a linear operator.

The substitutional and permutation operators already discussed are, by definition, linear operators.

Vector spaces invariant under an operator.

A vector space S will be said to be invariant under an operator A if $X \in S$ implies $AX \in S$. For example, the three-dimensional sub-space spanned by the vectors $\alpha(1)\alpha(2)\beta(3)$, $\alpha(1)\beta(2)\alpha(3)$, $\beta(1)\alpha(2)\alpha(3)$ is invariant under the permutation operators on the product space. We should note that this definition of invariance does not imply that the set of vectors $\{AX\}$ spans S, where X is an arbitrary vector in S, unless, in addition, the reciprocal of A exists. In general, there exists an unlimited number of spaces invariant under a given operator, differing in nature and dimensionality.

Matrix representation of an operator.

If $\{u_r\}$ is a basis in an n-dimensional space S invariant under an operator A, then by hypothesis $(Au_r) \in S$ $(r = 1 \ldots n)$ and $Au_r = \Sigma_s u_s U_{sr}(A)$. The $n \times n$ array of coefficients $U_{sr}(A)$ $(s = 1 \ldots n, r = 1 \ldots n)$ corresponding to the operator A is called the matrix representative of A in S relative to the basis $\{u_r\}$. For example, referred to the basis

$$\left.\begin{matrix}\alpha(1)\alpha(2)\beta(3)\\ \alpha(1)\beta(2)\alpha(3)\\ \beta(1)\alpha(2)\alpha(3)\end{matrix}\right\} \quad \begin{matrix}\text{the operator } P \equiv \boxed{\begin{matrix}1 & 2 & 3\\ 2 & 3 & 1\end{matrix}}\\ \text{is represented by the}\\ \text{matrix}\end{matrix} \quad \left.\begin{pmatrix}0 & 1 & 0\\ 0 & 0 & 1\\ 1 & 0 & 0\end{pmatrix}\right\}.$$

As a second example, consider an operator on our three-dimensional position space which permutes the coordinates (x, y, z) of a vector in one of 3! ways and changes the sign of the coordinates in one of 2^3 ways. Each of the $2^3 \times 3! = 48$ operators so defined is represented by a 3×3 matrix such that the three coefficients in each row and column consist of two zeros and either $+1$ or -1. Now let $\{A\}$ denote the set of 48 operators so defined and let X denote *any* non-zero vector in our space, then it is evident that the set of vectors $\{AX\}$, $A \in \{A\}$ spans the position space. This is not true of the sub-set of 3! permutation operations on the coordinates of X, for the vector $(1, 1, 1)$ is invariant under such permutations. We have here simple examples of relations between a set of operators and a vector space invariant under the set, which we shall discuss in § 4.

Given the matrix representative $\| U_{rs}(A) \|$ of a *linear* operator A relative to some basis $\{u_r\}$ in a space S, we can determine the effect of A on any arbitrary vector $X = \Sigma_s u_s X_s$ in S, for the linear property implies

$$AX = A \sum_s u_s X_s = \sum_s (Au_s) X_s = \sum_{r,s} u_r U_{rs}(A) X_s.$$

The explicit form of the matrix $\| U_{rs}(A) \|$ depends upon the basis chosen in S. For example, let $\{v_j\}$ denote an alternative basis in S, where

$$v_j = \sum_r u_r P_{rj}, \quad u_s = \sum_k v_k Q_{ks},$$

then if A is a linear operator,

$$Av_j = A \sum_r u_r P_{rj} = \sum_r (Au_r) P_{rj} = \sum_{r,s} u_s U_{sr}(A) P_{rj}$$

$$= \sum_k v_k \sum_{r,s} Q_{ks} U_{sr}(A) P_{rj};$$

so if $\| V_{kj}(A) \|$ denotes the matrix representative of A with respect to the basis $\{v_j\}$ in S, we have the relations

$$V_{kj}(A) = \sum_{r,s} Q_{ks} U_{sr}(A) P_{rj}, \qquad \begin{matrix} k = 1 \dots n, \\ j = 1 \dots n, \end{matrix}$$

between the matrix representatives of A corresponding to the alternative bases $\{u_r\}$ and $\{v_k\}$.

We may note here that $\chi(A)$, the sum of the diagonal elements (spur) of the matrix representative of a linear operator, is independent of the bases in S, for $\Sigma_j P_{rj} Q_{js} = \delta_{rs}$ from the basis nature of the sets $\{u_r\}$ and $\{v_j\}$, and therefore

$$\sum_j V_{jj}(A) = \sum_{j,r,s} Q_{js} U_{sr}(A) P_{rj} = \sum_{r,s} U_{sr}(A) \sum_j P_{rj} Q_{js} = \sum_r U_{rr}(A).$$

It follows that the spur $\chi(A)$ characterizes the relation between the operator A and the space S. In general, $\chi(A)$ is different, for different spaces invariant under A.

As a further demonstration of the importance of the linear property, we consider a space S invariant under a pair of linear operators A and B, and determine the matrix representatives of the products AB and BA (in general, $AB \neq BA$). It follows at once from the invariance of S and the linear properties of A and B, that

$$A(Bu_t) = A\left[\sum_s u_s U_{st}(B)\right] = \sum_r u_r \sum_s U_{rs}(A)\, U_{st}(B);$$

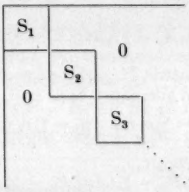
that is,

$$U_{rt}(AB) = \sum_s U_{rs}(A)\, U_{st}(B).$$

It is a simple generalization to consider a matrix representation of a set of operators. This is the subject of § 6. It is essential to remember that a set of operators possesses an infinite number of different matrix representations, depending both on the vector spaces invariant under the set and the bases chosen therein.

Direct sum of matrix representatives.

Let us suppose that a space S can be decomposed into a sum of non-overlapping sub-spaces $S_1, S_2, \dots$ (that is S_p, S_q $(p \neq q)$ contain no vectors in common), each invariant under an operator A, then we can choose and order a basis in S such that the basis vectors fall in one or other of these sub-spaces. We say that such a basis displays the decomposition of S into the sum of sub-spaces $S_1, S_2, \dots$. The matrix representative of A referred to such a basis will be of the block form



that is, the matrix representative of A in S is the direct sum of the matrix representatives of A in the component sub-spaces.

Direct product of matrix representatives.

If $S' \times S''$ is the product space constructed from spaces S', basis $\{u_p\}$ and S'', basis $\{v_r\}$, each invariant under an operator A, then we define the result of operating with a *linear* operator A on the vector $X = \Sigma_{p,r}(u_p v_r) X_{pr}$ in $S' \times S''$ to be the vector

$$AX = \sum_{p,r}(Au_p)(Av_r) X_{pr}.$$

It follows that the matrix representative of A in $S' \times S''$ relative to the basis $\{u_p v_r\}$ is given by the relations

$$A(u_p v_r) = (Au_p)(Av_r) = \sum_{q,s} u_q v_s U_{qp}(A) U_{sr}(A);$$

that is, the matrix representative of A in $S' \times S''$ is the direct product of the matrix representatives of A in S' and S''.

It is evident that this method of construction can be extended to the multiplication together of any number of invariant spaces and matrix representations.

Eigenvectors.

Given a vector space S invariant under an operator A, that is, $X \in S$ implies $AX \in S$, we can speculate as to whether it is possible to choose X such that the vector AX is some scalar multiple of X, i.e. $AX = aX$. In fact, if S is finite-dimensioned, it can be proved by exhaustion that at least one solution certainly exists. Such a vector X is called an eigenvector of A belonging to the eigenvalue a. An eigenvalue is said to be degenerate if two or more linearly independent eigenvectors exist corresponding to the given eigenvalue. If an operator A possesses as many linearly independent eigenvectors in S as the dimensionality of S, then it is obvious that these eigenvectors can be taken as a basis in S. Referred to this special basis, the matrix representative of the operator A is a diagonal matrix; that is, only the elements on the diagonal are non-zero.

Given a matrix representation of A, say $\| U_{rs}(A) \|$, we can find the eigenvalues of A by invoking the elementary theory of homogeneous linear equations. The eigenvalue equation $(A-a)X = 0$ is represented in matrix form by the set of equations

$$\sum_s [U_{rs}(A) - \delta_{rs} a] X_s = 0.$$

It follows that the eigenvalues are the roots of the determinantal equation $|U_{rs}(A) - \delta_{rs} a| = 0$.

A simple geometrical example of the foregoing is obtained if we consider A to be the reflection operator in our three-dimensional position space, corresponding to the mirror plane containing the z-axis, which interchanges the x- and y-axes. The matrix representative of this operator referred to these axes is

$$\begin{vmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{vmatrix}$$ while the determinantal equation for the eigenvalues is $$\begin{vmatrix} -a & 1 & 0 \\ 1 & -a & 0 \\ 0 & 0 & (1-a) \end{vmatrix},$$

that is, $(1-a)(a^2-1) = 0$. The root $a = 1$ is two-fold degenerate. If we choose as our basis the eigenvectors

$$(1\ 1\ 0), \quad (1\ -1\ 0), \quad (0\ 0\ 1),$$

the matrix representative takes the diagonal form

$$\begin{vmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{vmatrix}.$$

If the operator A in the equation $AX = aX$ is a linear operator, then the eigenvector X is not completely defined, for then $A(cX) = a(cX)$ where c is an arbitrary complex number. It follows that we can always normalize the eigenvector, that is, choose $(X.X) = 1$.

If a space S is invariant under a set of operators $\{A\}$, it may be possible to find vectors in S which are simultaneously eigenvectors of two or more operators in the set $\{A\}$, that is, vectors X such that $AX = aX, BX = bX, \ldots$. This is a familiar procedure in quantum mechanics.

The algebra of linear operators.

Let S denote a vector space invariant under a set of linear operators $\{I, A, B, C, \ldots\}$. For completeness we have introduced the unit operator I, which as its name implies, is such that $IX = X$ for all $X \in S$. By reference to this invariant vector space, we can define a non-commutative algebra of our operators over the field of complex numbers ; that is, we impose the conditions

$$(A+B)X = AX + BX = (B+A)X,$$

$$(aA)X = (Aa)X = a(AX),$$

$$\left.\begin{aligned}(AB)X &= A(BX)\\ (BA)X &= B(AX)\end{aligned}\right\} \text{ in general, } \quad A(BX) \neq B(AX),$$

$$(AB)CX = (ABC)X = A(BC)X,$$

$$(aA + bB)(cC + dD)X = ac(AC)X + ad(AD)X + bc(BC)X + bd(BD)X,$$

where X is an arbitrary vector in S and a, b, c, d any complex numbers. The operators A and B are said to commute if $(AB)X = (BA)X$, all $X \in S$. If an operator relation is true for all $X \in S$, we shall omit the vector X from the relation.

The process of division is not defined for our operators, for in general, the operator equation $AB = C$, where A and C are given operators, does not possess a solution B. A solution does exist if corresponding to the operator A, there exists an operator P such that $PA = I$ the unit operator ; for then, by the associative axiom, $B = PC$. This raises the question of the existence and definition of the reciprocal or inverse of an operator.

Reciprocal of an operator.

Given that an operator A exists, it does not follow that there exist operators P and Q, the left and right reciprocals of A respectively, such that $PA = I = AQ$; and even if P and Q exist, that $P = Q$. We shall only concern ourselves with operators such that if P and Q exist, then $P = Q$, and we shall write $P = A^{-1} = Q$. We should note here the identity $(A^{-1}B^{-1}) = (BA)^{-1}$; that is, the reciprocal of a product of operators exists, if the individual operators each possess a reciprocal.

The corresponding assertion is not true in general of a linear combination of a collection of operators; that is $(aA+bB)^{-1}$ cannot be assumed to exist without discussion.

Hermitian adjoint of an operator.

Given an operator A under which the unitary vector space S is invariant, we define an operator $A^\dagger$, called the Hermitian adjoint of A, by the condition

$$AX.Y = X.A^\dagger Y \text{ for all } X \text{ and } Y \text{ in S.}$$

This definition implies $(AB)^\dagger = B^\dagger A^\dagger$.

Unitary operators.

A linear operator A which possesses a reciprocal A^{-1} is said to be a unitary operator if $A^\dagger = A^{-1}$; that is,

$$\left.\begin{aligned} AX.Y &= X.A^{-1}Y \\ A^{-1}Y.X &= Y.AX \end{aligned}\right\} \text{ for all } X \text{ and } Y \text{ in S.}$$

This definition implies

$$AX.AY = X.Y \text{ for all } X \text{ and } Y \text{ in S;}$$

that is, a unitary operator preserves scalar products.

For example, let $\{u_r\}$ denote an orthonormal basis, i.e. $(u_r.u_s) = \delta_{rs}$, in a unitary space S invariant under a unitary operator A, then $\{v_r\}$, where $v_r = Au_r$, $u_r = A^{-1}v_r$, is also an orthonormal basis in S, for $(v_r.v_s) = Au_r.Au_s = u_r.u_s = \delta_{rs}$.

The product of unitary operators A and B is unitary, for

$$(AB)X.Y = X.B^{-1}A^{-1}Y = X.(AB)^{-1}Y;$$

but a linear combination of unitary operators is not necessarily unitary. It follows from the definition of a unitary operator that the reciprocal A^{-1} of a unitary operator A is unitary.

Hermitian operator.

A linear operator A is said to be Hermitian or self-adjoint if $A = A^\dagger$. The eigenvalues of a Hermitian operator are real, for given $A = A^\dagger$ and $AX = aX$, we note that

$$a(X.X) = (X.AX) = (AX.X) = a^*(X.X).$$

The eigenvectors of a Hermitian operator belonging to different eigenvalues are orthogonal; for $AX = aX$, $AY = bY$, $a \neq b$ implies

$$a(Y.X) = (Y.AX) = (AY.X) = b(Y.X); \text{ that is } (Y.X) = 0.$$

It can be shown (by exhaustion) for a finite dimensional space S invariant under a Hermitian operator A, that the eigenvectors of A in S span S. It follows that these eigenvectors can be chosen to form an orthonormal basis in S, for we can always construct orthonormal bases in the sub-spaces spanned by degenerate eigenvectors. In quantum mechanics it is necessary to *postulate* that the eigenvectors of a Hermitian operator form a complete basis in an infinite dimensioned space, because it is not possible to prove completeness except in special cases.

The most important Hermitian operator associated with a system is usually the Hamiltonian H, that is, the operator representing the energy of the system. The fact that we can build up a Hamiltonian operator from elementary Hermitian operators stems from the fact that a linear combination of Hermitian operators over the *real* numbers is Hermitian ; for

$$\begin{aligned}(aA+bB)X.Y &= (aAX.Y)+(bBX.Y)\\ &= X.(a^*A+b^*B)Y.\end{aligned}$$

The product of Hermitian operators is not Hermitian unless the operators commute, for $(AB)' = B'A' = BA$.

§ 4. Sets of Operators

In the last section we introduced the idea of a vector space invariant under an operator A. We now extend our ideas to the concept of a space S invariant under a set of operators $\{A\}$. This generalization has far-reaching consequences, for it introduces the concept of irreducibility. We say that S is irreducible under the set $\{A\}$, if, given any non-zero vector X in S, the set of vectors $\{AX\}$ spans S, where A runs over all the members of $\{A\}$; in other words, S is irreducible under $\{A\}$, if S contains no proper sub-space S′, such that S′ is invariant under $\{A\}$. We note, however, that in general, S irreducible under $\{A\}$ does not imply that S is irreducible under a sub-set of operators in $\{A\}$. We shall postpone until § 6 the technical problem of determining whether a space S is irreducible under a given set of operators $\{A\}$.

As an example of a space irreducible under a set of operators, we can take S to be the space defined by the basis vectors

$$u_1 = \alpha\alpha\beta+\omega\alpha\beta\alpha+\omega^2\beta\alpha\alpha,$$

$$u_2 = \alpha\alpha\beta+\omega^2\alpha\beta\alpha+\omega\beta\alpha\alpha,$$

taken from the eight-dimensional product space discussed in an earlier example. As members of the set $\{A\}$, we take the two permutation operators with their corresponding matrix representatives

$$\begin{vmatrix}1&2&3\\2&1&3\end{vmatrix} = \begin{pmatrix}0&1\\1&0\end{pmatrix} \quad \text{and} \quad \begin{vmatrix}1&2&3\\1&3&2\end{vmatrix} = \begin{pmatrix}0&\omega^2\\\omega&0\end{pmatrix}.$$

We shall give a proof of irreducibility in § 6.

Schur's lemma.

One of the most important theorems, which stems at once from the concept of irreducibility, is known as Schur's lemma. We shall not prove the complete theorem at this point, since we do not require full generality here. Given a finite space S, irreducible under a set of operators $\{A\}$, and invariant under a linear operator P, the lemma asserts that if $PA = AP$ for all $A \in \{A\}$, then P is a scalar multiple of the unit operator I ; that is, $P = cI$ (where the number c may be zero).

Proof.

Given an arbitrary basis $\{u_r\}$ in the n-dimensional space S, where

$$Au_r = \sum_s u_s U_{sr}(A), \quad \text{and} \quad PA = AP, \quad \text{all} \quad A \in \{A\},$$

consider the set of vectors $\{v_r = Pu_r\}$ generated by the operator P. These vectors span a sub-space of S, since S is invariant under P. We note that

$$Av_r = APu_r = PAu_r = \sum_s (Pu_s) U_{sr}(A) = \sum_s v_s U_{sr}(A);$$

that is, the sub-space S′ of S spanned by $\{v_r\}$ is invariant under $\{A\}$. But S is irreducible under A by hypothesis, therefore either $S' \equiv 0$ or $S' \equiv S$. It follows that either $P \equiv 0$, or the n vectors v_r are linearly independent, since they span the n-dimensional space S. In the latter case the determinant of the matrix $\| U_{rs}(P) \|$ cannot be zero, for $\det U_{rs}(P) = 0$ would imply that we could find a non-trivial solution to the equation $\Sigma_r v_r X_r = 0$, which would imply that the vectors $\{v_r\}$ were linearly dependent, in contradiction to hypothesis. Now given $AP = PA$, it follows that $A(P - cI) = (P - cI)A$ for an arbitrary number c. Therefore either $(P - cI) = 0$, or $\det(P - cI) \neq 0$ for arbitrary c. But the latter hypothesis is untenable, since we can deliberately choose c to be a solution of the determinantal equation $\det(P - cI) = 0$. Therefore $P = cI$.

As a demonstration of this theorem, consider our example of a two-dimensional space irreducible under the given pair of permutation operators. Given the hypothesis $AP = PA$ for all $A \in \{A\}$, that is $\Sigma_s U_{rs}(A) U_{sl}(P) = \Sigma_s U_{rs}(P) U_{sl}(A)$, the reader will find (by explicit calculation) that the only 2×2 matrix that commutes with the matrices

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 & \omega^2 \\ \omega & 0 \end{pmatrix} \quad \text{is of the form} \quad \begin{pmatrix} c & 0 \\ 0 & c \end{pmatrix}.$$

Decomposition of an invariant space.

The concept of irreducibility at once raises the question: can a space S, invariant under a set of operators $\{A\}$, be decomposed into a sum of non-overlapping sub-spaces $S', S'', \ldots$, each invariant and irreducible under $\{A\}$? The answer is yes—if certain conditions, which we now consider, are satisfied. It is important to note that we shall only give sufficient conditions; a discussion of necessary conditions would be too long and difficult here. We prove that if $\{A\}$ is a set of *unitary* operators, then complete decomposition of a finite space, invariant under $\{A\}$, is always possible.

Theorem on complete reducibility.

The theorem is proved by demonstrating, given a sub-space $S' \subset S$, S′ and S invariant under $\{A\}$, that we can always choose a sub-space S″ complementary to S′, that is, $S = S' + S''$ but S′ and S″ contain no non-zero vector in common, such that S″ is itself invariant under $\{A\}$, for the assertion of complete reducibility then follows by exhaustion, since S is finite.

Our demonstration consists of choosing the sub-space S″ to be an orthogonal complementary sub-space relative to S′, and then proving that the sub-space S″

so chosen is invariant under $\{A\}$. It is important to note that we demonstrate a *sufficient* condition for complete reducibility—the mutual orthogonality of S′ and S″ is not a *necessary* condition for complete reducibility, as we shall demonstrate by example in § 6. In order to prove that S″ so chosen is invariant under $\{A\}$, that is $Y \in S'', A \in \{A\}$, implies $AY \in S''$, we note that it follows from this choice of S″ that if $Y \in S$ and $X.Y = 0$ all $X \in S'$ then $Y \in S''$; for suppose $Y = Y_1 + Y_2$, where $Y_1 \in S'$ and $Y_2 \in S''$, then the condition $(X.Y) = 0$ for all $X \in S'$, together with the mutual orthogonality of S′ and S″, implies in particular $0 = (Y_1.Y) = (Y_1.Y_1)$, and therefore $Y_1 = 0$. Now the unitary property of A implies $AX.AY = X.Y$, so that the relation $X.Y = 0$ above implies $AX.AY = 0$. But $AX \in S'$ and $AY \in S$ since S′ and S are invariant under A, therefore $AY \in S''$; for the vector AX is an arbitrary vector in S′.

Having demonstrated that the unitary property of the operators is *sufficient* to ensure the complete reducibility of a vector space with respect to a set of operators, we now note that the unitary property is *not* necessary, for it is evident that:

(*a*) a space S invariant and irreducible under a set of unitary operators $\{A\}$ is automatically invariant and irreducible under an arbitrary linear combination of the operators contained in $\{A\}$;

(*b*) an operator which consists of an arbitrary linear combination of unitary operators is not in general itself unitary.

Groups.

The full beauty of the concept of a space S, irreducible under a set of operators, appears if the set is a group of unitary operators. Repeating our definition given in the Introduction, a set of operators $G \equiv \{A\}$ is said to form a group if three conditions are satisfied:

(1) The unit operator $I \in G$.

(2) $A \in G$ implies $A^{-1} \in G$.

(3) $\left.\begin{matrix} A \in G \\ B \in G \end{matrix}\right\}$ implies $\begin{matrix} (AB) \in G, \\ (BA) \in G. \end{matrix}$

If $G \equiv \{A\}$ contains a finite number of elements, G is called a finite group.

The group property does not imply the unitary property; for example, the operators of the Lorentz group, encountered in the special theory of relativity, are in general non-unitary, for the space–time continuum is not a unitary space. In solid-state physics, on the other hand, we are almost entirely concerned with finite groups of operators, and it will be shown later that the finite property implies the unitary property.

A group whose elements can be labelled with parameters that run over continuous ranges is called a continuous group. The group of all rigid rotations in our three-dimensional position space is a continuous group of unitary operators. The Lorentz group is a continuous group. As an example of a finite group we can take the set of six permutations on three symbols noted in our example of permutation operators, or the set of 24 orientation operations which rotate a cube into a cube.

The structure of a group is determined by its multiplication relations. Two groups of operators may be entirely different in their field of application and yet *isomorphic*; that is, possess the same abstract multiplication table. For example,

(*a*) our group of six permutations on three symbols is isomorphic with the group of geometrical operations generated by two mirror planes inclined at 120° to each other;

(*b*) the group of 24 orientation operations on a cube is isomorphic with the permutation group on four symbols.

It is for this reason that we speak of *abstract* groups, and consider groups of operators as realizations of abstract groups.

Homomorphism.

The relationship of isomorphism between two groups is a particular example of the more general relationship of homomorphism. A simple example of homomorphism is the relationship between the abstract group of two elements I and J where $J^2 = I$ on the one hand, and a group of permutations on the other. We note that all permutations are either even or odd, with the multiplication rule odd × odd = even, odd × even = even × odd = odd, which is identical with the multiplication structure of the two-group $G \equiv \{I, J\}$; in other words, we can map all the elements of a permutation group on the two-group. Generalizing this idea, a group G' is said to be a homomorphic image of a larger group G, if the elements of the larger group can be arranged in sets, such that if we disregard the differences between the elements of G in the same set, we abstract the image group G'. We shall find many examples of such homomorphisms when we come to the subject of matrix representations of groups in § 6.

Direct-product groups.

Many of the groups with which we shall be concerned are in fact products of two or more groups, say $G', G'', \ldots$, where the elements of each of the groups G' commute with all the elements of the other groups $G'' \ldots$. The groups $G', G'', \ldots$ may have elements in common; for example, the unit element is certainly common. A group G with this structure is said to be a direct-product group, and we write $G = G' \times G'' \ldots$. We can see at once, if S is a space invariant under G, and irreducible under any one of the groups G', that Schur's lemma has some very severe comments on such a situation. We consider these comments in § 6.

Group algebras.

In practice, the algebra of a finite group, which we now define, is a more useful concept than the group itself, just as the field of rational numbers is more useful than the set of integers. This concept is defined as follows.

Given a finite group G containing N linear operators g_n ($n = 1 \ldots N$), consider the infinite set $\mathscr{A}$ of linear operators generated from G by adding to the set all linear combinations of the elements g_n of G over the complex numbers; that is, the general operator A of $\mathscr{A}$ is the linear combination $A = \Sigma_n g_n a_n$. We can regard $\mathscr{A}$ as an N-dimensional vector space, for it is possible to prove by appealing to the theory of irreducible group representations that the relation $\Sigma_n g_n a_n = 0$

implies $a_n = 0$, all n. As an example we can construct a six-dimensional space from the group of six permutations on three symbols.

Such a vector space, called a group algebra, differs from the vector spaces so far encountered, in that we can multiply two vectors in $\mathscr{A}$ together to give a vector in $\mathscr{A}$; for given $A = \Sigma_n g_n a_n$ and $B = \Sigma_n g_n b_n$ in $\mathscr{A}$, the ordered product of A and B is

$$AB = \sum_m g_m a_m \sum_n g_n b_n = \sum_{m,n} (g_m g_n)(a_m b_n) = C \in \mathscr{A},$$

since $(g_m g_n) \in G$, for all m, n, from the group property. This property means that we can regard any given element of $\mathscr{A}$ either as a vector *in* $\mathscr{A}$ or as an operator *on* $\mathscr{A}$, the choice depending on the needs of the moment. We shall make use of this property when we consider the regular representation of a finite group in § 6.

Having defined such a vector space using the group elements as the basis vectors, we can change to any other convenient basis just as we are accustomed to do in the case of simple vector spaces. In § 7, where we shall describe the practical applications of group algebras, we shall choose particular linear combinations of the group elements as our basis vectors.

The dimensionality of irreducible spaces.

It is a remarkable fact that the irreducible vector spaces corresponding to a given group of operators, although of infinite variety in their origin, are severely constrained in their group-theoretical properties particularly in dimensionality. For example, although vector spaces irreducible under the symmetry group of the cube can originate in many different ways, their dimensionality is constrained to be either 1, 2 or 3—no other results are possible. The groups of permutation operations, which are relevant to the many-body problem and the Pauli principle, are a second example of these constraints; for among the great variety of vector spaces irreducible under the symmetric groups, as they are called, only two kinds of one-dimensional space invariant under permutation operators can exist. In one case the basis vector X is unchanged under all permutation operations P; i.e. $PX = X$. In the other case, $PX = X$ if the permutation is even; but $PX = -X$ if the permutation P is odd.

In general, a given vector space S, irreducible under a set $\{A\}$, is reducible under a sub-set of $\{A\}$. On the other hand, if S invariant under a set $\{A\}$, is irreducible under a sub-set of $\{A\}$, then S is irreducible under $\{A\}$. For example, if S is invariant and irreducible under a set of unitary operators $\{A\}$, then S is invariant and irreducible under the algebra $\mathscr{A}$ generated from the set $\{A\}$.

We shall develop the idea of a vector space irreducible under a group of operators in § 6 on matrix representations, but first we consider the relevance of the idea to quantum mechanics.

§ 5. Symmetry in Quantum Mechanics

In this section we discuss the connection between group theory and quantum mechanics. Since quantum mechanics is concerned with the description of the physical world, a discussion of this nature necessarily involves our assumptions about the nature of the real world. In fairness, the reader should be warned that alternative opinions are held on the relationship between group theory and quantum mechanics and the opinion given here is not universal.

Degeneracy induced by symmetry.

The central postulate of quantum mechanics is that corresponding to an isolated physical system there exists a vector space S_0 called the state space of the system and a Hermitian operator H called the Hamiltonian of the system, such that the eigenvectors of H span S_0, the eigenvalues of H being the energy levels of the system. These eigenvalues are commonly degenerate. If the system possesses symmetry of some kind this is manifested in the existence of symmetry operators that commute with the Hamiltonian. The connection between the existence of such operators and the occurrence of degenerate eigenvalues of H is easily demonstrated; for let the operator A commute with H, i.e. $HA = AH$, then if $X(E)$ is an eigenvector of H belonging to the eigenvalue E, we have $H(AX) = A(HX) = A(EX) = E(AX)$, assuming that $AE = EA$; so if the vector (AX) is linearly independent of $X(E)$, we have degeneracy induced by the symmetry operator A. We now go further and show that to each eigenvalue E of H there corresponds a sub-space $S(E)$ of S_0, spanned by the eigenvectors belonging to E, which is invariant under the group of unitary symmetry operators that commute with H, the symmetry group of the Hamiltonian.

The symmetry group of a Hamiltonian operator.

The symmetry group of a Hamiltonian H is the complete set G of *unitary* operators that commute with H. To show that this set G is a group, we note that the definition of G is such that

(1) G contains the unit operator I, for $IH = HI$,

(2) $A \in G$ implies $A^{-1} \in G$, for $A^{-1}H = HA^{-1}$ and A^{-1} is unitary,

(3) $\left.\begin{matrix} A \in G \\ B \in G \end{matrix}\right\}$ implies $\left\{\begin{matrix} AB \in G \\ BA \in G \end{matrix}\right.$ for $\left.\begin{matrix} ABH = HAB \\ BAH = HBA \end{matrix}\right\}$ and AB and BA are unitary.

The symmetry properties of a quantum-mechanical system, that give rise to the existence of the symmetry group of the system, may take a variety of forms. The most easily recognized form is orientational symmetry, e.g. cubic symmetry, while the symmetry associated with the Pauli principle and spin is more subtle in form.

In practice it is impossible to be certain that we have found the complete set of symmetry operators belonging to a given system. There always exists the possibility, for example, that the symmetry group we have discovered is a subgroup of the complete symmetry group of the system. We shall return to this question when we come to our postulate of irreducibility.

It is often the case that the symmetry group of a Hamiltonian is a direct product of two or more groups of operators. For example, in an n-electron model, the symmetry group is a direct product group, one of whose factors is the permutation group on the n symbols labelling the electrons in the model.

Given a Hamiltonian operator H of the form $H = \Sigma_r H_r$ where the terms H_r represent various physical features of the system, for example, kinetic energy, potential energy, spin–orbit coupling, interaction with external electromagnetic fields, we are faced with the fact that each term H_r defines a symmetry group G_r, our task being to find the symmetry group G of H. If the nature of the symmetry operators is the same for all the elements H_r, it is evident that the group G is the

set of operators common to the sets G_r. If, however, the operators are not all of the same nature, we may appear to be in difficulty. For example, as we explain in Part 2, § 2, the presence of spin–orbit coupling terms makes it necessary to define symmetry operators which consist of operators on the spin-space of the system, together with simultaneous substitutional operations on the orbital space of the system. The solution to this dilemma is to replace each group G_r by an abstract group $\mathscr{G}_r$ of which G_r is a homomorphic image, such that the groups $\mathscr{G}_r$ are all sub-groups of some abstract group G_0. The symmetry group of $H = \Sigma_r H_r$ is the set of elements of G_0 common to the abstract sub-groups $\mathscr{G}_r$.

The eigenvector sub-spaces $S(E)$.

Let $S(E)$ denote the sub-space of the state space S_0 spanned by the eigenvectors of H belonging to the eigenvalue E. To show that each sub-space $S(E)$ is invariant under the symmetry group G of H, we note that given $X \in S(E)$ then $HX = EX$ and $H(AX) = E(AX)$, for all $A \in G$; that is, $X \in S(E)$ implies $(AX) \in S(E)$ for all $A \in G$.

The postulate of irreducibility.

We now introduce a non-trivial relation between group theory and quantum mechanics by requiring, except in certain well-defined cases, that these eigenvector sub-spaces $S(E)$ shall be *irreducible* under the *complete* symmetry group of H. We do this because experience has shown that this postulate leads to agreement between the predictions of quantum mechanics and experiment. This postulate implies, if experiment reveals an order of degeneracy exceeding that required by the recognized set of symmetry operators, that either we have made a mistake, or the set of symmetry operators is incomplete—in other words, *accidental* degeneracy does not occur in the real world. We shall only apply the adjective accidental to spurious degeneracies that occur in over-simplified models.

One of the simplest examples of these ideas is provided by a Hamiltonian H which is invariant under spatial inversion in a point. Let J denote this operation, then the eigenvectors X of H can be classified into two types, even and odd, with respect to the operator J according as $JX = \pm X$ respectively. If J and the unit operator I are the only symmetry operators, then according to the degeneracy postulate above, all the sub-spaces $S(E)$ belonging to the given H are one-dimensional, for the group (I, J) has no other irreducible representations.

The exceptions to our postulate on degeneracy, namely that the sub-spaces $S(E)$ of S_0 shall be irreducible under G, arise when the Hamiltonian of a system is such that there exist *non-linear* symmetry operators that carry eigenvectors in one irreducible eigenvector sub-space $S(E)$ into another $S(E')$, and thus $E = E'$. A simple example of this occurs in the case of a real Hamiltonian H, i.e. $H^* = H$, for the operation of taking the complex conjugate commutes with H, that is, $(HX)^* = H^* X^* = HX^*$. Additional degeneracy arises, then, if $X \in S(E)$ but $X^* \notin S(E)$. A more sophisticated example occurs in the case of the Dirac Hamiltonian for an electron, if there is no external magnetic field present, for in this case there exists a non-linear operator R (the Wigner time-reversal operator) which commutes with H and induces extra degeneracy if $RX(E) \notin S(E)$. (This is the explanation of Kramers degeneracy in spectroscopy.)

Perturbation splitting of degeneracy.

An important application of our degeneracy postulate is to the problem of the effect of a perturbation of the Hamiltonian of a system on the degeneracy of the energy levels. We have noted that a vector space irreducible under a set of operators $\{A\}$ is, in general, reducible under a sub-set of $\{A\}$. Let G_0 be the symmetry group of a Hamiltonian H_0 and let G_1 be the symmetry group of the perturbed Hamiltonian $H_1 = H_0 + \alpha V$ where the parameter $0 \leqslant \alpha \leqslant 1$. The group G_1 will be a sub-set (sub-group) of G_0. Now given a multi-dimensional sub-space $S(E)$ of eigenvectors of H_0 irreducible under G_0, we can, by the theorem on complete reducibility, decompose $S(E)$ into a sum of sub-spaces, each irreducible under the sub-group G_1 of G_0. It now follows from the degeneracy postulate that as the parameter α increases from zero, these sub-spaces irreducible under G_1 go over into the irreducible eigenvector spaces of $H_1 = H_0 + \alpha V$. We can say, therefore, that if the original space $S(E)$ is irreducible under the sub-group G_1 then the perturbation will not affect the degeneracy of the energy level E, but if $S(E)$ is reducible under G_1 then the degeneracy will be split in a manner corresponding to the complete reducibility of $S(E)$ with respect to the sub-group G_1. The splitting of atomic energy levels by applied electric or magnetic fields is an example of this phenomenon.

The symmetry group of models.

In order to make a problem tractable, it is often necessary in quantum mechanics to simplify the physical Hamiltonian by leaving out terms which we know should be present on physical grounds, for example, electron–electron interactions, spin–orbit coupling, etc. The symmetry group of the resulting simplified Hamiltonian will then contain the symmetry group of the physical Hamiltonian as a sub-group. We have here a group-theoretical situation closely related to the situation in the foregoing discussion on perturbation splitting of degeneracy. A common procedure is to work with a model whose symmetry group is a direct-product group. For example, in a one-electron problem, if we neglect the spin–orbit coupling terms, the symmetry group becomes the direct product of the spin-rotation group with the orbital symmetry group.

The Pauli principle and Schur's lemma.

Let us consider a quantum-mechanical system of n equivalent particles. The Hamiltonian of the system will commute with the $n!$ permutations on the labels attached to the particles for the purposes of calculation; in other words, the symmetry group of the Hamiltonian certainly contains the group of permutations Π_n on n symbols. Suppose, for simplicity, that no other symmetry operators exist, then according to the postulate of irreducibility, we should expect to find the energy eigenvector sub-spaces $S(E)$ irreducible under the group Π_n. Now suppose that a given sub-space $S(E)$, irreducible under Π_n, is multi-dimensional, and let us consider how we could, in principle, distinguish by experiment between the degenerate states corresponding to $S(E)$. Let P denote an operator representing an observable quantity distinguishing the states of our system, and let us propose to use the eigenvalues of P to distinguish between the degenerate states corresponding to $S(E)$. Our efforts are doomed to failure, however, for by Schur's

lemma, the eigenvalues of P associated with $S(E)$ are all the same, for $S(E)$ is irreducible under Π_n, and P commutes with all the elements of Π_n, since P is symmetric in the particle labels. It follows, in so far as quantum mechanics is relevant to the real world, that we would not expect to find this kind of degeneracy in the real world. When we consider the different kinds of one-dimensional spaces invariant under Π_n, we find that there are only two kinds, the symmetric and the anti-symmetric; in other words, we would expect to find only bosons and fermions in the real world.

This section has been concerned with the deep-lying relationship between group theory and quantum mechanics. In contrast to this aspect of the subject, there is the strictly utilitarian use of group theory in quantum mechanical computational problems, in particular, the calculation of eigenvectors and eigenvalues by approximate methods. As we might expect, the intimate relation between the symmetry group and the invariant eigenvector spaces manifests itself in computational schemes for finding eigenvalues and eigenvectors. In order to exploit these group-theoretical ideas to our computational advantage we require a suitable calculus. For this reason we now develop the calculus of the irreducible matrix representations of groups.

§ 6. Matrix Representations

The object in the preceding sections of this article was to introduce the concept of irreducibility, and then to relate this concept to quantum mechanics. Looking back, we can see that the ideas involved are relatively few and simple, most of our effort being expended on the careful framing of definitions and relations. Now, when we turn to the technical application of group theory to the solution of problems, we shall find that the new ideas, although still simple, are relatively numerous, compared with what has gone before. This section is, in fact, a condensation of a large part of the book-work of group theory, and the reader should be fair to himself and not expect to digest this concentrated fare with the same speed with which, it is hoped, he digested the first five sections of this article.

A second point we make here is that the techniques to be developed in this and § 7 are those required in solid-state theory proper, where the symmetry groups are finite groups. We do not describe the very elegant techniques developed and exploited by, amongst others, Elliott, Judd and Runciman (1957), in order to calculate the energy levels of ions in a host crystal. These are essentially problems in atomic spectroscopy, involving the theory of angular momentum (Edmonds 1957, Fano and Racah 1959).

Matrix representations.

We observed in § 3, given an n-dimensional space S invariant under a set of operators $\{A\}$, that corresponding to any chosen basis $\{e_r\}$ in S, there exists an $n \times n$ matrix representation of $\{A\}$, defined by the relations $Ae_s = \Sigma_r e_r U_{rs}(A)$, where the set $\{U(A)\}$ of $n \times n$ matrices is called the matrix representation of $\{A\}$, relative to the basis $\{e_r\}$, afforded by the space S. If the invariant space is one-dimensional, the matrix representatives are numbers (1×1 matrices).

This concept of a matrix representation of a set of operators is important for our purpose because the most characteristic features of vector spaces invariant

under a set of operators $\{A\}$ are the corresponding matrix representations of $\{A\}$. When we come to the subject of group characters we shall see that given certain information about the matrix representations of a finite group G afforded by a space S, we can at once determine whether S is reducible or irreducible under G, and if reducible, the decomposition of S into a sum of sub-spaces each irreducible under G. This is the information we require, we may recall, for example, to determine the splitting of energy levels under a perturbation.

It is conventional to use the expression ' irreducible matrix representation ' to denote a matrix representation afforded by a representation space irreducible under a given set of operators $\{A\}$. Such an irreducible representation will be denoted by the symbol $\{U^{\lambda}(A)\}$ where the use of a superscript will enable us to distinguish between essentially different irreducible representations—a problem which we shall consider when we come to the subject of equivalent representation spaces. It is important to realize that the irreducible matrix representations of a group are properties of the underlying abstract group, for these matrices can be calculated from the multiplication table of the abstract group alone. In practice it is rarely necessary to calculate *ab initio* the irreducible matrix representations of a group of operators, for once the underlying abstract group has been recognized, this information can be found in textbooks.

Isomorphic and homomorphic representations.

In our definition in § 3 of the matrix representatives of operators, we noted that the matrix representative of a product of *linear* operators is the matrix product of the representatives; that is, if $AB = C$, then $U(A)U(B) = U(C)$. If the correspondence between $\{A\}$ and the image $\{U(A)\}$ is not faithful, that is, one-to-one (isomorphic), we call $\{U(A)\}$ a homomorphic or many–one representation of $\{A\}$. We may, in other words, have two spaces S and S′, each invariant under $\{A\}$, but in S the members $A, B, \ldots$ of a sub-set of $\{A\}$ are all represented by the same matrix, although in S′, the matrix representatives of the members of the sub-set are all distinct. The extreme example of this situation occurs if S is one-dimensional, and such that for every vector $X \in$ S and $A \in \{A\}$, we have $AX = X$. This representation is known as the unit representation of $\{A\}$.

It is not uncommon for a group of operators G to contain an infinite number of distinguishable elements in one context, while a given homomorphic representation is a finite group G'; that is, each element of G' is the image of an infinite number of elements in G. This situation occurs, for example, in the case of matrix representations of the space-groups of infinite crystal lattices.

The regular representation of a finite group.

A particularly important faithful representation of a finite group is the regular or permutation representation, which is defined as follows.

Let S denote an N-dimensional space invariant under a finite group G of order N, such that S contains a non-zero vector X with the property that the N vectors $AX, A \in G$ are linearly independent. Taking these vectors as a basis in S, consider the matrix representation of G in S with respect to the basis $\{AX\}$. It is evident from the group multiplication relations that the rows and columns of each $N \times N$ matrix representative contain $+1$ in some position and $(N-1)$ zeroes

elsewhere. This set of matrices is called the regular representation of G. The precise form of each matrix follows at once from the group multiplication table.

As an example, let G denote the group of three cyclic permutations on the coordinates (x, y, z) of a vector in our three-dimensional spaces, and take X to be the vector $(1, 0, 0)$, then the other two vectors of the set are $(0, 1, 0)$ and $(0, 0, 1)$. The corresponding 3×3 matrix representatives are

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \quad \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

The regular representation of a finite group is a reducible representation, with the remarkable property that when this representation is expressed as a sum of irreducible representations, it is found that each irreducible representation occurs as many times as its dimensionality. This we shall prove when we come to the section on group characters.

In our example the group of three cyclic permutations has three one-dimensional irreducible representations, which combine in the form

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 & 0 \\ 0 & \omega & 0 \\ 0 & 0 & \omega^2 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 & 0 \\ 0 & \omega^2 & 0 \\ 0 & 0 & \omega \end{pmatrix} \qquad \omega = -\frac{1}{2} + i\frac{\sqrt{3}}{2}$$

which is the representation we should have obtained if, instead of our basis e_x, e_y, e_z, we had used the basis of vectors :

$$e_1 = (e_x + e_y + e_z), \quad e_2 = (e_x + \omega e_y + \omega^2 e_z), \quad e_3 = (e_x + \omega^2 e_y + \omega e_z).$$

We may note here for future reference in § 7 that the regular representation of a finite group is manifested if we ask for the matrix representatives of the group elements in a group algebra, using the group elements as basis vectors in the vector space afforded by the group algebra.

Reciprocal representatives.

It is important to note that in a matrix representation we have the useful relations $U^{-1}(A) = U(A^{-1})$; for by definition, we have the matrix relations

$$U^{-1}(A)\,U(A) = U(A)\,U^{-1}(A) = U(I) = U(A^{-1})\,U(A) = U(A)\,U(A^{-1}),$$

and so

$$U^{-1}(A) = U(A^{-1})\,U(A)\,U^{-1}(A) = U(A^{-1}).$$

Unitary and orthogonal matrices.

We shall be specially concerned with unitary and orthogonal matrices. A unitary matrix U is a matrix whose reciprocal U^{-1} is equal to the complex conjugate of its transpose ; that is, $U^{-1} = U^{\dagger}$, where $U^{\dagger} = (\tilde{U})^*$ is called the Hermitian adjoint of U. A real unitary matrix O is called an orthogonal matrix. In this case $O^{-1} = \tilde{O}$.

Equivalent representations.

We now consider in what way matrix representations of a given set of linear operators may differ. Representations of different dimensions are obviously different, so let us consider representations of the same dimensions. Consider then, two spaces S_1 and S_2, each invariant under $\{A\}$ and of equal dimensions, say n. In other respects S_1 and S_2 may be entirely different. Let $\{U_{pq}(A)\}$ denote the $n \times n$ matrix representation of $\{A\}$ in S_1 relative to a basis $\{u_p\}$, $p = 1 \ldots n$. Let $\{V_{jk}(A)\}$ denote the $n \times n$ matrix representation of $\{A\}$ in S_2 relative to a given basis $\{v_j\}$, $j = 1 \ldots n$. The question is: can we choose a basis in S_2 such that the corresponding matrix representation of $\{A\}$ is identical to $\{U_{pq}(A)\}$ in S_1? In order to discuss this question, let $\{V_{pq}(A)\}$ denote the representation of $\{A\}$ in S_2 relative to an arbitrary basis $\{e_p\}$, related to the given basis $\{v_j\}$ in S_2 by the relations

$$e_p = \sum_j v_j P_{jp}, \quad v_k = \sum_q e_q P_{qk}^{-1}.$$

We can display these ideas schematically in the form

$$\begin{array}{lll} \text{Basis } \{u_p\} & & \text{given basis } \{v_j\},\ \text{repr. } \{V_{jk}(A)\} \\ & S_1 \longleftrightarrow S_2 & \\ \text{Repr. } \{U_{pq}(A)\} & & \text{arbitrary basis } \{e_p\},\ \text{repr. } \{V_{pq}(A)\}. \end{array}$$

The relationship between the representations $\{V_{jk}(A)\}$ and $\{V_{pq}(A)\}$ follows from the equations

$$Ae_p = \sum_j (Av_j) P_{jp} = \sum_{j,k} v_k V_{kj}(A) P_{jp} = \sum_q e_q \sum_{j,k} P_{qk}^{-1} V_{kj}(A) P_{jp};$$

that is

$$V_{pq}(A) = \sum_{j,k} P_{pk}^{-1} V_{kj}(A) P_{jq}, \quad \text{for all} \quad A \in \{A\}.$$

Our question can now be expressed in the form: is it possible to choose the matrix $\|P_{jp}\|$ such that the resulting matrix representation $\{V_{pq}(A)\}$ of $\{A\}$ in S_2 is identical to the given representation $\{U_{pq}(A)\}$ of $\{A\}$ in S_1; that is $V_{pq}(A) = U_{pq}(A)$, for all A in $\{A\}$? If this transformation P exists, the spaces S_1 and S_2 are said to be equivalent with respect to $\{A\}$, and the given representations $\{U_{pq}(A)\}$ and $\{V_{jk}(A)\}$ are said to be equivalent matrix representations, related by the similarity transformation

$$U_{pq}(A) = \sum_{j,k} P_{pk}^{-1} V_{kj}(A) P_{jq},$$

where the matrix $\|P_{jp}\|$ is independent of A in $\{A\}$. If the spaces S_1 and S_2 are not equivalent, they are said to be inequivalent representation spaces with respect to $\{A\}$.

Adjoint representations.

We note here that if $\{U(A)\}$ is any matrix representation of a set of linear operators $\{A\}$, and $\{\zeta(A)\}$ is a one-dimensional representation of $\{A\}$, that is, the representatives $\zeta(A)$ are complex numbers, then the set of matrices $\{\zeta(A)U(A)\}$ is a representation of $\{A\}$, since $\zeta(A)U(A) = U(A)\zeta(A)$. The representation so generated may or may not be equivalent to $\{U(A)\}$. For example, the group Π_n of permutations on n symbols has a one-dimensional representation in which every even permutation is represented by $+1$ and every odd permutation by -1.

Given any representation of Π_n, we can construct another representation of Π_n. These are called adjoint (or associated) representations. If the representations are equivalent, they are said to be self-adjoint.

Isomorphic bases.

We shall make frequent use of isomorphic bases, that is bases in equivalent representation spaces so chosen that the corresponding matrix representations are identical. We have an example of isomorphic bases in our eight-dimensional product space invariant under the group of permutations on three symbols, for the bases $\{\alpha(1)\alpha(2)\beta(3),\ \alpha(1)\beta(2)\alpha(3),\ \beta(1)\alpha(2)\alpha(3)\}$ and $\{\beta(1)\beta(2)\alpha(3),\ \beta(1)\alpha(2)\beta(3),\ \alpha(1)\beta(2)\beta(3)\}$ are isomorphic bases in the two invariant three-dimensional sub-spaces.

Products of representations.

Given two spaces S′ and S″, each invariant under a set of linear operators $\{A\}$, we can consider the matrix representations of $\{A\}$ afforded by the product space $S' \times S''$. Let $\{u_r\}$ denote a basis in S′, where $Au_s = \Sigma_r u_r U_{rs}(A)$; and let $\{v_p\}$ denote a basis in S″, where $Av_q = \Sigma_p v_p V_{pq}(A)$; then we can take $\{u_r v_p\}$ to be a set of basis vectors in $S' \times S''$. The matrix representative of A relative to this basis in $S' \times S''$ is exhibited by the equation

$$A(u_s v_q) = \sum_{r,p} (u_r v_p) U_{rs}(A) V_{pq}(A).$$

This representation of $\{A\}$ in $S' \times S''$ is called the product of the representations $\{U(A)\}$ and $\{V(A)\}$. Recalling our definition in § 3 of the spur of an operator, the spur of A relative to $S' \times S''$ is

$$\chi(A) = \sum_{s,q} U_{ss}(A) V_{qq}(A) = \Big[\sum_s U_{ss}(A)\Big]\Big[\sum_q V_{qq}(A)\Big],$$
$$= \chi'(A)\chi''(A),$$

where $\chi'(A)$ and $\chi''(A)$ are the spurs of A relative to S′ and S″, respectively. It is evident that this construction of representations can be extended to the product of any number of representations.

Characters.

In our definition of the spur $\chi(A)$ of an operator A relative to a space S, we noted that $\chi(A)$ is independent of the basis chosen in S. It follows that the set of spurs $\{\chi(A)\}$, $A \in \{A\}$, called the character of S relative to $\{A\}$, is a distinguishing feature of the relation between $\{A\}$ and S, which does not require explicit reference to a particular basis. By definition, equivalent representation spaces possess identical characters; that is, identity of characters is a necessary condition for equivalence. We shall see later that the condition is also sufficient.

We may note here that the character of the regular representation of a finite group of order N is such that $\chi(I) = N$ and $\chi(A) = 0$, $A \neq I$.

The advantage of using the language of characters is not only that we can discuss the decomposition of representation spaces and matrix representations without reference to particular sets of basis vectors, for we can go further and omit reference to the spaces and matrices themselves, referring only to the characters

that define them sufficiently for our purpose. When we speak, as we shall presently, of the decomposition of a reducible character into a sum of irreducible characters, we shall imply a decomposition of the corresponding representation space *and* the decomposition of the reducible matrix representations without specifying the choice of basis.

Irreducible group characters.

The characters of representation spaces, irreducible under a group of unitary operators $G \equiv \{A\}$, are important for our purpose because

(*a*) the irreducible group characters $\{\chi^i(A)\}, \{\chi^j(A)\}, \ldots$ are characteristic of the underlying abstract group and can be calculated from the group multiplication table alone;

(*b*) our theorem on complete reducibility implies that the character $\{\chi(A)\}$ of a space S, reducible under $\{A\}$, is the sum of irreducible characters $\chi^i(A)\}$, $\{\chi^j(A)\}, \ldots$, in the sense that $\chi(A) = \Sigma_i c_i \chi^i(A)$, for all $A \in G$, where the integer c_i is the number of times irreducible sub-spaces with character $\{\chi^i(A)\}$ appear in the decomposition of S;

(*c*) the irreducible characters corresponding to a given group possess remarkable orthogonality properties, which are particularly simple to describe and use in the case of finite groups. If $\chi^i \equiv \{\chi^i(A)\}$ is an irreducible character of a finite group $\{A\}$ of order N, then

$$\sum_A \chi^i(A)\chi^i(A^{-1}) = N;$$

while if χ^i and χ^j are two different irreducible characters of $\{A\}$, then

$$\sum_A \chi^i(A)\chi^j(A^{-1}) = 0, \quad i \neq j.$$

These orthogonality relations for the irreducible characters of finite groups, as they are called, are a special case of more general orthogonality relations, which we shall invoke later. These relations are in fact a direct consequence of Schur's lemma, but the derivation, being lengthy, is relegated to the Appendix. It follows at once from these relations that $\Sigma_A \chi(A)\chi(A^{-1}) = N$ is a *necessary* condition that a given character $\{\chi(A)\}$ should be irreducible. That this condition is also *sufficient* for irreducibility follows from the observation that if $\{\chi(A)\}$ is reducible, that is $\chi(A) = \Sigma_j c_j \chi^j(A)$, then $\Sigma_A \chi(A)\chi(A^{-1}) > N$, for

$$\sum_A \chi(A)\chi(A^{-1}) = \sum_A \sum_i c_i \chi^i(A) \sum_j c_j \chi^j(A^{-1}) = N \sum_j c_j^2 > N,$$

since the assumed reducibility of $\chi(A)$ implies that $\Sigma_j c_j^2 > 1$.

The orthogonality relations are often written in the form

$$\sum_A \chi^i(A)\chi^{*j}(A) = \delta_{ij} N,$$

for, as we shall see later, in the case of finite groups, $\chi^j(A^{-1}) = \chi^{*j}(A)$ the complex conjugate of $\chi^j(A)$.

It follows at once from the orthogonality relations, if a given representation $\{U^j(A)\}$ of a finite group $\{A\}$ is irreducible, that an adjoint (associated) representation $\{\zeta(A)\,U^j(A)\}$, where $\{\zeta(A)\}$ is a one-dimensional representation of $\{A\}$, is

irreducible; for the spur of $\zeta(A)U^j(A)$ is $\zeta(A)\chi^j(A)$ and $\zeta(A)\zeta(A^{-1}) = 1$, therefore

$$\sum_A \zeta(A)\chi^j(A)\zeta(A^{-1})\chi^j(A^{-1}) = \sum_A [\zeta(A)\zeta(A^{-1})]\chi^j(A)\chi^j(A^{-1}) = N.$$

Using the orthogonality relations, we can deduce a formula for the number of times c_i a given irreducible character $\{\chi^i(A)\}$ occurs in the decomposition of a reducible character $\{\chi(A)\}$, for

$$\frac{1}{N}\sum_A \chi(A)\chi^i(A^{-1}) = \frac{1}{N}\sum_j c_j \sum_A \chi^j(A)\chi^i(A^{-1}) = c_i.$$

We have thus demonstrated that the decomposition of a reducible character is unique. Furthermore, it is now evident that identity of character is a sufficient as well as a necessary condition for the equivalence of two representation spaces, whether they be reducible or irreducible, for an irreducible character cannot be orthogonal to itself.

Using the formula above for the coefficients c_i in the decomposition $\chi(A) = \sum_i c_i\chi^i(A)$, we can prove the assertion that in the decomposition of the character of the regular representation of a finite group G, each irreducible character $\{\chi^i(A)\}$ occurs as many times as the dimensionality of the representation, that is $c_i = n_i$, where n_i = dimensionality of the ith irreducible representation of G. The proof follows at once from the fact that in the regular representation the only non-zero spur belongs to the identity operator I, and $\chi(I) = N$ the order of G,

$$c_i = \frac{1}{N}\chi(I)\chi^i(I) = \chi^i(I) = n_i.$$

It follows at once that

$$\sum_i n_i^2 = N \quad \text{since} \quad \sum_i c_i n_i = N.$$

We shall return to this remarkable relation when we consider group algebra in more detail in § 7.

We can use our formula for the coefficients c_i to calculate the number of energy levels into which a given degenerate level is split by a perturbation. Let $\{\chi(A)\}$ denote the character of the eigenvector sub-space $S(E)$, irreducible under the symmetry group G_0 of the unperturbed Hamiltonian H_0, and let $\{\chi^i(R)\}$ denote the ith irreducible character belonging to the finite sub-group G_1, the symmetry group of the perturbed Hamiltonian, then according to the degeneracy postulate, in the absence of non-linear symmetry operators, the energy level E will be split into $\Sigma_j c_j$ levels where

$$\sum_j c_j = \frac{1}{N}\sum_j \sum_R \chi(R)\chi^j(R^{-1}).$$

Note that the validity of this result does not require that the group G_0 be finite.

Character analysis of product representations.

We have noted earlier that the spur $\chi(A)$ of an operator A relative to a product space $S = S_n \times S_m \times \ldots$, built up from spaces $S_n, S_m, \ldots$ invariant under $\{A\}$, is the product of the spurs $\chi_n(A), \chi_m(A), \ldots$ of A relative to the component spaces; that is, $\chi(A) = \chi_n(A) \times \chi_m(A) \ldots$. If the set $\{A\}$ is a finite group, we can use our

group character formulae to decompose the character of S relative to $\{A\}$. We can write formally $\chi = \Sigma_j c_j \chi^j$, where

$$c_j = \frac{1}{N} \sum_A [\chi_n(A) \chi_m(A) \ldots] \chi^j(A^{-1}).$$

This type of analysis is important in many-electron problems. (The Clebsch–Gordan decomposition formula is an example of this idea applied to a many-particle system with spherical symmetry.)

Classes.

Given any element A of a group, consider the set of distinct group elements of the form $P^{-1}AP$, generated as P runs over all the elements of the group. This set of equivalent elements is called a class of the group. Equivalent elements, A and B say, possess the same spur in a given representation space, for $B = P^{-1}AP$ implies

$$\chi(B) = \sum_r U_{rr}(B) = \sum_{r,s,t} U_{rs}(P^{-1}) U_{st}(A) U_{tr}(P) = \sum_s U_{ss}(A) = \chi(A).$$

It follows that it is sufficient to give the spur of a class in published tables of group characters.

Using the concept of a class, we now derive, using our old friend Schur's lemma, a set of equations for the irreducible characters of a finite group, from which the characters may be calculated. Our first step is to note, given a class of equivalent operators in a group, that we can form the algebraic sum of these operators. We call the operator so formed a class sum or a class operator. In the case of finite groups, a class sum is an element of the group algebra. Confining ourselves now to finite groups, we note that the class operators of a finite group form a sub-algebra of the group algebra ; that is, the product of two class operators C_α, C_β can be expressed as a linear combination of the complete set of class operators. We have then

$$C_\alpha C_\beta = \sum_\gamma h_{\alpha\beta\gamma} C_\gamma.$$

The proof of this assertion is almost self-evident. We note that by definition, $P^{-1} C_\alpha P = C_\alpha$ for an arbitrary group element P, and therefore

$$P^{-1} C_\alpha C_\beta P = P^{-1} C_\alpha P P^{-1} C_\beta P = C_\alpha C_\beta.$$

It follows that if a given group element g occurs n times in the expansion of $C_\alpha C_\beta$ as a sum of group elements, then every element equivalent to g occurs n times, for if this were not so, we could find an element P such that $P^{-1} C_\alpha C_\beta P \neq C_\alpha C_\beta$.

The coefficients $h_{\alpha\beta\gamma}$ are found from the group multiplication table.

We now note, since C_α commutes with every element of the group, that it follows from Schur's lemma that the matrix representative of C_α in an irreducible matrix representation of the group algebra is a scalar multiple x_α of the unit matrix, where $r_\alpha \chi_\alpha^i = x_\alpha n_i$, n_i being the dimensionality of the irreducible space, r_α the number of group elements in the class α, and χ_α^i the representative spur for the class α. It follows that the matrix representation of the relation $C_\alpha C_\beta = \Sigma_\gamma h_{\alpha\beta\gamma} C_\gamma$ is $x_\alpha x_\beta = \Sigma_\gamma h_{\alpha\beta\gamma} x_\gamma$, which on substitution for x_α gives the set of equations for the irreducible characters

$$r_\alpha r_\beta \chi_\alpha^i \chi_\beta^i = n_i \sum_\gamma h_{\alpha\beta\gamma} r_\gamma \chi_\gamma^i.$$

In fact, the number of inequivalent irreducible characters is equal to the number of classes of a finite group, but the proof is too long to give here. The character of the unit representation is included in this total.

Irreducible representations of direct-product groups.

We consider here a certain type of irreducible representation of a direct-product group. Consider a space S, invariant under the direct-product $G = G' \times G''$, and simultaneously irreducible under G'; then S is irreducible under G, since G' is a sub-group of G. By definition of a direct product, the elements of G'' commute with all the elements of G'. It follows therefore, from Schur's lemma, that the matrix representatives in S, of the elements of G'', are all scalar multiples of the unit matrix. Moreover, these scalar multiples afford a one-dimensional representation of G''. It follows that only a limited class of irreducible representations of a direct-product group are to be found in such a situation.

If $G = G' \times G''$ is a finite group such that the only element common to G' and G'' is the unit element, we can, given all the irreducible representations of G' and G'', find all the irreducible representations of G as follows. Let S′ and S″ be spaces, invariant under G, where S′ is irreducible under G' and affords the unit representation of G'', while S″ is irreducible under G'' and affords the unit representation of G'. Now consider the representation of $G = G' \times G''$ afforded by the product space S′ × S″. Let $\{u_r\}$ denote a basis in S′ and $\{v_p\}$ a basis in S″, where for all g' and g'',

$$(g'g'')u_s = \sum_r u_r U_{rs}(g'), \quad (g'g'')v_q = \sum_p v_p V_{pq}(g''),$$

$\{U(g')\}$ being an irreducible representation of G', and $\{V(g'')\}$ an irreducible representation of G''. We have then, for the basis $\{u_s v_q\}$ in S,

$$(g'g'')u_s v_q = \sum_{r,p} u_r v_p U_{rs}(g') V_{pq}(g''),$$

$$\chi(g'g'') = \chi(g')\chi(g'').$$

Since g' and g'' are finite groups it follows at once from the properties of irreducible characters that this representation of $G = G' \times G''$ is irreducible; for

$$\sum_{g'g''} |\chi(g'g'')|^2 = \sum_{g'} |\chi(g')|^2 \sum_{g''} |\chi(g'')|^2,$$

$$= N' \times N'' = \text{order of } G,$$

since by hypothesis S′ is irreducible under G' and S″ is irreducible under G'', where G' and G'' possess only the unit element in common. Moreover, we can find all the irreducible representations of $G = G' \times G''$ by this means, for the total number to be found is equal to the number of classes of equivalent elements in G, and this is equal to the product of the number of classes in G' with the number of classes in G'', since G' and G'' commute and possess only the unit element in common. A simple example of this situation is that in which G'' contains only two elements, say I and J where $J^2 = I$. The group G'' then has only two irreducible representations, one being the unit or even representation of G'', while the other is the odd representation of G'' (in which J is represented by -1). The group $G = G' \times G''$ then has twice as many irreducible representations as G', half of these being

identical with the irreducible representations of G', while the other half are odd with respect to the operation J; that is, the product $(g'J)$ is represented by $-U(g')$.

We return to the question of the irreducible representation of direct-product groups at the end of this section.

Unitary representations and isomorphic bases.

Up to this point in our argument, the choice of a particular basis in a given vector space has been irrelevant, and we have emphasized this point by using the language of characters. From this point onwards, however, the choice of basis becomes important, because in order to exploit the properties of irreducible representation spaces, it is necessary that two conditions shall be satisfied by our choice of basis. The first constraint imposed is such that the matrix representatives of unitary operators are unitary matrices.

In order to arrange that the matrix representatives of unitary operators A on a given space S are unitary matrices, that is, $U^{-1}(A) = U^{\dagger}(A)$, we may choose an orthogonal basis $\{e_r\}$ in S, normalized such that $(e_r . e_s) = a\delta_{rs}$, where $a > 0$ is an arbitrary positive number, for then

$$\delta_{rs} = \frac{1}{a}(e_r . e_s) = \frac{1}{a}(Ae_r . Ae_s) \quad \text{since } A \text{ is unitary,}$$

$$= \frac{1}{a}\left\{\sum_p e_p U_{pr}(A)\right\} \cdot \left\{\sum_q e_q U_{qs}(A)\right\},$$

$$= \frac{1}{a}\sum_{p,q}(e_p . e_q) U_{pr}^*(A) U_{qs}(A) = \sum_p U_{pr}^*(A) U_{ps}(A);$$

that is,

$$U(I) = U^{\dagger}(A)U(A), \text{ and therefore } U^{-1}(A) = U^{\dagger}(A), \text{ since } U(I) = U^{-1}(A)U(A).$$

The fact that we do not need to normalize the basis in S to unity will be exploited in § 7.

It is important to note that we have demonstrated that the orthogonality condition $(e_r . e_s) = a\delta_{rs}$ on the basis $\{e_r\}$ in S is sufficient to ensure that the matrix representatives are unitary—the orthogonality condition is, in fact, necessary if S is irreducible under a finite group of operators. As an example of a reducible unitary representation, derived from a non-orthogonal basis, we may take the regular representation of a finite group. This is a real unitary representation but there is no requirement that the basis be orthogonal.

We turn aside here to outline a proof of the assertion made earlier that the irreducible characters of a finite group possess the property $\chi^j(A^{-1}) = \chi^{*j}(A)$. We note first that if A is a unitary operator on a space S, then $\chi^*(A) = \chi(A^{-1})$, for we have shown, if A is a unitary operator on a space S, that we can choose the basis $\{e_r\}$ in S such that the matrix representative $U(A)$ is unitary, i.e. $U^{\dagger}(A) = U^{-1}(A)$; but $U^{\dagger}(A) = U^{-1}(A) = U(A^{-1})$, therefore $\chi^*(A) = \chi(A^{-1})$ independent of the choice of basis in S. The assertion $\chi^j(A^{-1}) = \chi^{*j}(A)$ is certainly true therefore in the case of groups of unitary operators. Now every finite group of order N can be regarded as a sub-group of the permutation group Π_N on N symbols. This is evident from our discussion of the regular representation of a finite group. Furthermore we can always construct a vector space which exhibits Π_N as a group of

unitary operators. It follows that every finite group is isomorphic with a group of unitary operators; therefore, since the irreducible characters depend only on the structure of the abstract group, it follows that $\chi^j(A^{-1}) = \chi^{*j}(A)$ is true for all finite groups.

We return now to the second constraint which we shall impose on our choice of basis in a representation space. If in the decomposition of a space S reducible under a finite group G, we discover irreducible sub-spaces $S(p,i)$, $S(q,i)\ldots$ with the same character $\{\chi^i(A)\}$ say, we choose the orthogonal bases in these equivalent irreducible sub-spaces to be isomorphic. To exhibit this choice, we use the notation $u(p,r,i)$ to denote the rth basis vector in the n_i-dimensional sub-space $S(p,i)$ etc., and then by definition,

$$Au(p,r,i) = \sum_s u(p,s,i)\, U_{sr}{}^i(A), \quad r = 1 \ldots n_i,$$

that is, the operator A is to be represented by the same $n_i \times n_i$ matrix $U^i(A)$ in each of the sub-spaces $S(q,i)$. In addition, we partially relax the orthogonality condition on the basis in S, to the extent that we do not require the irreducible *equivalent* sub-spaces $S(p,i)$, $S(q,i)\ldots$ to be mutually orthogonal. Finally, we shall not require the basis vectors in the sub-space $S(p,i)$ to be normalized to the same 'length' as in the sub-space $S(q,i)$; that is, the normalization is to be arbitrary among the sets of isomorphic bases; for example we may have

$$u(p,r,i).u(p,s,i) = a\delta_{rs}, \quad a > 0$$

but

$$u(q,r,i).u(q,s,i) = b\delta_{rs}, \quad b > 0 \quad \text{with} \quad b \neq a.$$

This freedom of choice of normalization will be useful later.

A basis $\{u(p,r,i)\}$ chosen in this way has certain remarkable properties which we shall exploit in § 7. These properties can be summarized in the two relations

$$u(p,r,i).u(q,s,j) = 0 \quad \text{if} \quad i \neq j, \quad \text{and if} \quad r \neq s;$$

$$u(p,r,i).u(q,r,i) \quad \text{is independent of } r.$$

The derivation of these properties depends on the generalized orthogonality relations for finite groups which were foreshadowed in our discussion of irreducible group characters. These general relations, which we prove in the Appendix, can be stated in two parts:

(1) If $\{U_{rs}{}^i(A)\}$ and $\{U_{\alpha\beta}{}^j(A)\}$ are inequivalent irreducible representations of a finite group then

$$\sum_A U_{rs}{}^i(A)\, U_{\alpha\beta}{}^j(A^{-1}) = 0, \quad \text{for all } r,s,\alpha,\beta.$$

(2) For any given $n_i \times n_i$ irreducible matrix representation $(U_{rs}{}^i(A))$ of a finite group of order N,

$$\sum_A U_{pq}{}^i(A)\, U_{rs}{}^i(A^{-1}) = \frac{N}{n_i}\delta_{ps}\,\delta_{qr}.$$

It is usual to find these relations quoted in the combined form

$$\sum_A U_{rs}{}^i(A)\, U_{\alpha\beta}{}^j(A^{-1}) = \frac{N}{n_i}\delta_{ij}\,\delta_{r\beta}\,\delta_{s\alpha}.$$

The orthogonality relations for the irreducible characters of a finite group follow at once, for

$$\sum_A \chi^i(A)\chi^j(A^{-1}) = \sum_{r,\alpha}\sum_A U_{rr}{}^i(A)\,U_{\alpha\alpha}{}^j(A^{-1}) = N\delta_{ij}\frac{1}{n_i}\sum_{r,\alpha}\delta_{r\alpha} = N\delta_{ij}.$$

Given the general relations, we proceed as follows,

$$u(p,\alpha,j).u(q,r,i) = \frac{1}{N}\sum_A [Au(p,\alpha,j)].[Au(q,r,i)]$$

since the operators A are unitary,

$$= \sum_{\beta,s}[u(p,\beta,j).u(q,s,i)]\frac{1}{N}\sum_A U^*{}_{\beta\alpha}{}^j(A)\,U_{sr}{}^i(A),$$

$$= \sum_{\beta,s}[u(p,\beta,j).u(q,s,i)]\frac{1}{N}\sum_A U_{sr}{}^i(A)\,U_{\alpha\beta}{}^j(A^{-1})$$

since the matrix representations have been chosen to be unitary,

$$= \frac{1}{n_i}\sum_s [u(p,s,i).u(q,s,i)]\,\delta_{ij}\,\delta_{\alpha r},$$

from the orthogonality relations, since the bases in equivalent spaces have been chosen to be isomorphic ;

where we note that $\Sigma_s[u(p,s,i).u(q,s,i)]$ is independent of r. We have set out this derivation in detail, in order to show how the properties of a basis of the form $\{u(p,r,i)\}$ depend on the definition of such a basis.

These innocent looking properties of a basis $\{u(p,r,i)\}$ derive their significance for our purpose from the fact that if H is a linear operator such that $AH = HA$, $A \in G$, then $A\{Hu(p,r,i)\} = H\{Au(p,r,i)\}$, since $AH = HA$,

$$= \sum_s \{Hu(p,s,i)\}\,U_{sr}{}^i(A), \quad \text{since } H \text{ is linear,}$$

and therefore repeating the foregoing arguments,

$$u(p,r,i).Hu(q,s,j) = 0 \quad \text{if } i \neq j, \text{ and if } r \neq s\,;$$

$$u(p,r,i).Hu(q,r,i) \quad \text{is independent of } r.$$

We may note here :

(*a*) the relation $u(p,r,i).u(q,s,j) = 0$ if $i \neq j$, implies that inequivalent irreducible sub-spaces $S(p,i)$, $S(q,j)$ $(i \neq j)$ are necessarily mutually orthogonal, in the sense that $X \in S(p,i)$, $Y \in S(q,j)$ $(j \neq i)$, implies $(X.Y) = 0$;

(*b*) a scalar product of the form $X.HY$ is called, in quantum mechanics, a matrix element of H taken between the vectors X and Y. This kind of matrix element can always be defined, whereas our definition of a matrix representative of an operator A, relative to a space S, requires that S be invariant under A. The two definitions coincide if S is invariant under A, and we choose an orthonormal basis $\{e_r\}$ in S ; for given $Ae_l = \Sigma e_s U_{sl}(A)$, and $(e_r.e_s) = \delta_{rs}$, we have

$$(e_r.Ae_s) = \sum_l (e_r.e_l)\,U_{ls}(A) = U_{rs}(A).$$

Decomposition of a space relative to a direct-product group.

The symmetry group of the Hamiltonian of a system of n similar particles, say electrons, is the direct product of the group of permutations Π_n on the particle labels, with the space-symmetry group G of the system, for these two kinds of symmetry operation commute with each other. Consider then a space S invariant under the direct product $G \times \Pi_n$, where we assume G to be finite. Let us begin the decomposition of S by decomposing S with respect to Π_n ; that is, we choose a basis in S of the form $\{u(p, r, \lambda)\}$, where λ labels the irreducible representations of Π_n, and r numbers the basis in the sub-spaces $S(p, \lambda)$ irreducible under Π_n. Now consider the p_λ-dimensional sub-space $S(r, \lambda)$ of S, spanned by the set of p_λ vectors $u(p, r, \lambda)$, $(p = 1 \ldots p_\lambda)$, corresponding to given r and λ, where p_λ is the number of times the irreducible character λ occurs in the decomposition of the character of S relative to Π_n.

This space $S(r, \lambda)$, basis $\{u(p, r, \lambda)\}$ $(p = 1 \ldots p_\lambda)$, is invariant under the group G, for consider the effect of operating upon a vector $u(p, r, \lambda)$ in $S(r, \lambda)$ with $g \in G$. Since the space S is invariant under g, we can write

$$gu(p, r, \lambda) = \sum_{q, s, \mu} u(q, s, \mu)\, C(g : q, s, \mu : p, r, \lambda),$$

where the coefficients $C(g : q, s, \mu : p, r, \lambda)$ are the coordinates of the vector $gu(p, r, \lambda)$ with respect to our basis in S. We now note, from the properties of this basis, that

$$C(g : q, s, \mu : p, r, \lambda) = u(q, s, \mu) . gu(p, r, \lambda) = 0$$

$$\text{if} \quad s \neq r \quad \text{and if} \quad \mu \neq \lambda ;$$

for $g \in G$ commutes with every element in Π_n. It follows that the space $S(r, \lambda)$ is invariant under G. Moreover, the coefficient $C(g : q, r, \lambda : p, r, \lambda)$ is independent of the label r. It follows, in particular, that the character of the space $S(r, \lambda)$ relative to G is independent of r, for the spur of $g \in G$, relative to $S(r, \lambda)$ is

$$\sum_p C(g : p, r, \lambda : p, r, \lambda).$$

We can now choose isomorphic bases in the sub-spaces $S(r, \lambda)$ of the form $\{u(q : r, \lambda : \alpha, j)\}$, where j labels the irreducible representations of G and α numbers the basis in the sub-spaces $S(q : r, \lambda : j)$. We then have by definition

$$gu(q : r, \lambda : \alpha, j) = \sum_\beta u(q : r, \lambda : \beta, j)\, U_{\beta\alpha}{}^j(g).$$

The set of basis vectors corresponding to given labels (q, λ, j) is such, given $P \in \Pi_n$ and $g \in G$, that

$$(Pg)\, u(q : r, \lambda : \alpha, j) = \sum_{s, \beta} u(q : s, \lambda : \beta, j)\, U_{\beta\alpha}{}^j(g)\, U_{sr}{}^\lambda(P) ;$$

that is, the representation of $G \times \Pi_n$ in this basis is the direct product of the irreducible representation $\{U^\lambda(P)\}$ of Π_n with the irreducible representation $\{U^j(g)\}$ of G. Now G and Π_n are finite groups with only the unit element in common ; it follows therefore, from our earlier discussions of this kind of representation, that our representation of $G \times \Pi_n$ in the sub-space $S(q, \lambda, j)$ is irreducible.

§7. The Group-theoretical Simplification of Eigenvalue Calculations

Many problems in solid state physics involve the calculation of eigenvalues and eigenvectors of Hermitian operators, but in only very special cases is it possible to carry out exact calculations, and we normally have to resort to approximate solutions, the limit on the accuracy of such solutions depending on the computational effort available. If the operator in question possesses a symmetry group as defined in § 5, we can use our group theory to deploy the available computational effort to maximum advantage.

We may note here that it is sometimes argued that the advent of the large fast computing machine has made group-theoretical assistance unnecessary. This, of course, is fallacious. In solid-state physics, the problem of calculating the properties of a system of interacting particles demands a computational effort, which will always exceed the capacity of the largest and fastest computer, for we continually increase the number of particles in our models, in order to approach a realistic description of a solid. For this central problem, group-theoretical techniques will always enhance the value of a given computer.

We introduce the group-theoretical method by a discussion of a hypothetical problem which in principle could be solved exactly in a finite time. Such a problem might be the determination of the normal modes of a vibrating molecule. By using such an example to start with, we can consider the difficulties in stages.

Consider then an Hermitian operator H with finite symmetry group of unitary operators $G \equiv \{A\}$. Let S denote an n-dimensional space invariant under H—this is the condition that we can rarely satisfy in practice. Furthermore, let S be invariant under G—this can always be arranged since G is finite. We require the eigenvalues E, and eigenvectors X, of H in S.

Let us ignore the existence of G for the moment, and proceed by a sledge-hammer method; namely, given a basis $\{e_r\}$ in S we can expand an eigenvector X in S in the form $X = \Sigma_r e_r X_r$. Substituting this expansion in the eigenvalue equation $HX - EX = 0$, we obtain the equation

$$\sum_r [He_r - Ee_r] X_r = 0.$$

We next construct a set of n linear homogeneous equations in the coordinates X_r, by forming the scalar product of the vector $\Sigma_r [He_r - Ee_r] X_r$ with each of the basis vectors e_s. The result is the set of n linear homogeneous equations

$$\sum_r [(e_s . He_r) - E(e_s . e_r)] X_r = 0, \quad s = 1 \ldots n;$$

with the $n \times n$ determinantal condition on E,

$$\det |(e_s . He_r) - E(e_s . e_r)| = 0.$$

The roots of this secular equation, as it is called, are the eigenvalues of the operator H belonging to the vector space S.

Now let us exploit the symmetry of the problem. Our first step is to calculate the character of the space S with respect to G. We can then choose the basis in S to be of the form $\{u(p, r, i)\}$. We order this basis as follows—all vectors with

the same r, i labels are collected together, c_i in number, where c_i is the number of times the ith irreducible character of G occurs in the decomposition of the character of S. Then by virtue of the property

$$\left.\begin{aligned} u(p,r,i).u(q,s,j) &= 0 \\ u(p,r,i).Hu(q,s,j) &= 0 \end{aligned}\right\} \quad \text{if } i \neq j, \text{ and if } r \neq s,$$

the secular determinant

$$|\{u(p,r,i).Hu(q,s,j)\} - E[u(p,r,i).u(q,s,j)]| = 0$$

factorizes into factor determinants, each factor being labelled with a pair of labels r and i, the order of the factor determinant being $c_i \times c_i$. Furthermore, by virtue of the property

$$\left.\begin{aligned} u(p,r,i).u(q,r,i) \\ u(p,r,i).Hu(q,r,i) \end{aligned}\right\} \quad \text{independent of } r,$$

corresponding to each label i, there will appear n_i identical factor determinants, where n_i is the dimensionality of the ith irreducible representation of G, so manifesting the n_i-fold degeneracy of the eigenvalues corresponding to eigenvector spaces with the irreducible character $\chi^i(A)\}$ of G. Moreover, if for some reason we are interested only in eigenvalues associated with eigenvectors of a particular symmetry, it is sufficient to construct the corresponding factor determinant.

We postpone the description of a special technique for calculating a basis of the form $\{u(p,r,i)\}$ until we have discussed approximate calculations, the technique being common to both types of calculation.

Approximate eigenvalues and eigenvectors.

In practice we are usually faced with the fact that the only space available to us which is invariant under the operator H of interest is of infinite dimensionality, so giving rise to an infinite secular determinant. Suppose however that H is of the form $H = H_0 + H'$ where we know the eigenvectors of the operator H_0. Furthermore, suppose we know on physical grounds that H' is such that the eigenvectors of H can be approximated by a linear combination over a finite set of eigenvectors of H_0, say $\{e_r\}$, $r = 1 \ldots n$. In this case we determine the 'best' combination $X = \Sigma_r e_r X_r$, either by a variational argument, or by the simpler argument that leads to the same equations for the coefficients X_r, namely that the best values for the coefficients X_r are given by the solutions of the n equations

$$\sum_r [(e_s.He_r) - W(e_s.e_r)]\, X_r = 0, \quad s = 1 \ldots n,$$

on the grounds that these equations go over into the exact equations without change of form as $n \to \infty$. The approximate eigenvalues W are then given by the solutions to the $n \times n$ secular equation

$$|(e_s.He_r) - W(e_s.e_r)| = 0.$$

The symmetry arguments go through in exactly the same way as in an exact calculation. If necessary we extend the space S spanned by the vectors $\{e_r\}$ to ensure

that S is invariant under the symmetry group G of H. We then choose a basis of the form $\{u(p,r,i)\}$ in S where p and i are determined by the character of S with respect to G. Ordering the basis as before, the secular equation decomposes into factors of the form

$$|u(p,r,i).Hu(q,r,i) - W^i u(p,r,i).u(q,r,i)| = 0, \quad \begin{matrix} p = 1 \dots c_i, \\ q = 1 \dots c_i, \end{matrix}$$

there being n_i identical factors corresponding to the ith irreducible character occurring in the decomposition of the character $\{\chi(A)\}$ of S.

Technique for constructing a basis $\{u(p,r,i)\}$.

The first step in constructing a basis of the form $\{u(p,r,i)\}$ in a given space S invariant under a finite group of unitary operators $G \equiv \{A\}$ is to determine the character of S with respect to G. In principle we can do this by calculating the diagonal elements $U_{rr}(A)$ of the matrix representatives of one operator A from each class of G, referred to some convenient basis in S. In practice it may be worth using an indirect method which we describe in the Appendix. For the present we shall assume that $\{\chi(A)\}$ of S is known and thus the integers

$$c_j = N^{-1} \sum_A \chi(A) \chi^j(A^{-1}).$$

Suppose then that we know that c_j irreducible sub-spaces $S(p,j)$, $p = 1 \dots c_j$, with character $\{\chi^j(A)\}$ will occur in the decomposition of the space S. Let $\{U_{rs}^j(A)\}$ denote a known irreducible unitary matrix representation of G with the character $\{\chi^j(A)\}$. Corresponding to this representation, we construct the n_j^2 elements e_{rs}^j of the group algebra of G, defined by the relations

$$e_{rs}^j = \frac{n_j}{N} \sum_A A U_{sr}^j(A^{-1}) = \frac{n_j}{N} \sum_A A U^*_{rs}{}^j(A), \quad \begin{matrix} r = 1 \dots n_j, \\ s = 1 \dots n_j, \end{matrix}$$

$\{U^j(A)\}$ being unitary. Since $\Sigma_j n_j^2 = N$, the order of G, from our discussion of the regular representation, we can construct a set of N such operators, corresponding to any complete set of irreducible representations of G. These operators e_{rt}^j possess the property

$$A e_{rt}^j = \sum_s e_{st}^j U_{sr}^j(A), \quad \text{for all } A \in G;$$

for

$$A e_{rt}^j = \frac{n_j}{N} \sum_B A B U_{br}^j(B^{-1}),$$

so on writing $AB = C, B^{-1} = C^{-1}A, U_{br}^j(B^{-1}) = \Sigma_s U_{bs}^j(C^{-1}) U_{sr}^j(A)$, and summing over C in place of B, since A is constant, we find

$$A e_{rt}^j = \frac{n_j}{N} \sum_C C \sum_s U_{bs}^j(C^{-1}) U_{sr}^j(A) = \sum_s e_{st}^j U_{sr}^j(A).$$

Now let ψ denote a non-zero vector in our space S and consider the relation

$$A(e_{rt}^j \psi) = (A e_{rt}^j)\psi = \sum_s (e_{st}^j \psi) U_{sr}^j(A), \quad A \in G.$$

This relation implies that if the vectors $(e_{sl}^{j}\psi)$ are non-zero, then $(e_{rl}^{j}\psi)$ is of the form $u(p,r,j)$. We shall show presently that if we can find ψ and e_{rl}^{j} such that $(e_{rl}^{j}\psi) \neq 0$, then $(e_{sl}^{j}\psi) \neq 0$, $s = 1 \ldots n_j$. It follows that if we can satisfy this condition, we can at once construct one of the c_j sub-spaces $S(p,j)$ in S. In order to find all c_j sub-spaces, we make use of the lemma (which we shall prove) that given some other vector, say $(e_{su}^{j}\phi)$, then either $(e_{su}^{j}\phi) = a(e_{sl}^{j}\psi)$ where a is a number independent of s, or the n_j vectors $(e_{su}^{j}\phi)$, $s = 1 \ldots n_j$, are linearly independent of the set $(e_{sl}^{j}\psi)$. It follows, if there are only two sub-spaces $S(p,j)$, $p = 1, 2$ (a common situation), that we can choose linearly independent vectors by inspection.

The most complete example of this technique occurs in the decomposition of a space S, affording the regular representation of a finite group $G \equiv \{A\}$. Let ψ denote a vector in S, such that the N vectors $(A\psi)$, $A \in G$, are linearly independent, then we know at once that $(e_{rs}^{j}\psi) \neq 0$, for all (j,r,s); for $(e_{rs}^{j}\psi) = 0$ implies that the vectors $\{A\psi\}$ are linearly dependent, contrary to hypothesis. Moreover $(e_{rs}^{j}\psi)$ and $(e_{rl}^{j}\psi)$ are linearly independent if $s \neq t$ for $(e_{rs}^{j}\psi) = a(e_{rl}^{j}\psi)$ again implies that the vectors $\{A\psi\}$ are linearly dependent. It follows that the $N = \Sigma_j n_j^2$ vectors $\{e_{rs}^{j}\psi\}$ form a basis $\{u(s,r,j)\}$ in S, which displays the complete decomposition of the regular representation of G, each irreducible representation of G occurring as many times as its dimensionality, as we found in § 6.

We summarize the decomposition procedure as follows:

(1) Calculate the character of S relative to G.

(2) Express the character of S as a sum of irreducible characters of G; that is, $\chi = \Sigma_j c_j \chi^j$.

(3) Corresponding to χ^j for which $c_j \neq 0$, construct one of the n_j^2 operators e_{rs}^{j} from a known unitary irreducible representation of G. Since the choice of the labels r, s is at our disposal, we first choose that combination for which as many as possible of the corresponding matrix elements in the irreducible representation are zero.

(4) Choose a convenient vector ψ in S, namely one for which it is relatively easy to calculate the vectors $(A\psi)$. Operate on ψ with the chosen operator e_{rs}^{j}.

(5) If $e_{rs}^{j}\psi = 0$, make a different choice of s. If the result is persistently zero, it means that, by chance, we have chosen a vector ψ which has no component in the irreducible sub-spaces of S, with character χ^j. We then make a different choice of ψ.

(6) Assuming that we have chosen e_{rs}^{j} and ψ such that $(e_{rs}^{j}\psi) \neq 0$, we now have to find a set of c_j linearly independent vectors of this form. If we vary our choice of the 'column' subscript, that is, the second subscript attached to e_{rs}^{j}, we may be fortunate and obtain c_j linearly independent vectors, using the same raw material ψ. If we find less than c_j independent vectors for a given ψ, we make a different choice of ψ and repeat the process, until we have satisfied our requirements.

The great advantage of using these projection operators, as they are called, is the fact that they automatically construct the isomorphic bases we require in our factorization procedure for secular determinants. The technique presupposes, of course, that irreducible unitary matrix representations of G are known at the outset.

In addition to the foregoing properties, these projection operators possess the algebraic properties :

$$e_{rs}^{i} e_{uv}^{j} = \delta_{ij}\,\delta_{su}\, e_{rv}^{i} \quad \text{(to be proved)},$$

$$(e_{rs}^{i})^{\dagger} = e_{sr}^{i}, \quad \text{since the matrices } U^{i}(A) \text{ and the group elements } A \text{ are unitary},$$

$$e_{rs}^{i} H = H e_{rs}^{i} \quad \text{since } AH = HA, \text{ for all } A \in G, \text{ by definition of } G.$$

It follows, given $u(p,r,i) = e_{rs}^{i}\psi$, $u(q,r,i) = e_{rt}^{i}\phi$, that

$$u(p,r,i).u(q,r,i) = (e_{rs}^{i}\psi).(e_{rt}^{i}\phi) = \psi.e_{sr}^{i}e_{rt}^{i}\phi = \psi.e_{st}^{i}\phi\,;$$

$$u(p,r,i).Hu(q,r,i) = (e_{rs}^{i}\psi).H(e_{rt}^{i}\phi) = \psi.e_{sr}^{i}He_{rt}^{i}\phi,$$
$$= \psi.He_{sr}^{i}e_{rt}^{i}\phi = \psi.He_{st}^{i}\phi,$$

so simplifying the calculation of the elements of the secular determinant

$$|[u(p,r,i).Hu(q,r,i)] - W^{i}[u(p,r,i).u(q,r,i)]| = 0.$$

It is important to note that these operators do not construct basis vectors normalized to unity, for $(e_{rs}^{i}\psi).(e_{rs}^{i}\psi) = \psi.e_{ss}^{i}\psi$, independent of r. This is why we left the normalization undefined in the definition of the vectors $u(p,r,i)$, apart from the constraint $[u(p,r,i).u(p,r,i)]$ independent of r.

If the irreducible character χ^{j} occurs once, and once only, in the decomposition of the character of the space S, that is, $c_j = 1$, we do not require an explicit irreducible representation of G, for the following reason. The corresponding root W^{j} of the secular equation is given formally by

$$(e_{rs}^{j}\psi).H(e_{rs}^{j}\psi) - W^{j}(e_{rs}^{j}\psi).(e_{rs}^{j}\psi) = 0.$$

This relation is true for all choices of the subscript s, for $(e_{rt}^{j}\psi) = a(e_{rs}^{j}\psi)$, since $c_j = 1$. It follows that

$$\sum_{s}[(e_{rs}^{j}\psi).H(e_{rs}^{j}\psi)] - W^{j}\sum_{s}[(e_{rs}^{j}\psi).(e_{rs}^{j}\psi)] = 0\,;$$

that is,
$$\psi.H\sum_{s}e_{ss}^{j}\psi - W^{j}\psi.\sum_{s}e_{ss}^{j}\psi = 0.$$

Now
$$\sum_{s}e_{ss}^{j} = \frac{n_j}{N}\sum_{A}A\sum_{s}U^{*}{}_{ss}^{j}(A) = \frac{n_j}{N}\sum_{A}A\chi^{*j}(A).$$

Therefore
$$W^{j} = \frac{\sum_{A}\chi^{*j}(A)(\psi.HA\psi)}{\sum_{A}\chi^{*j}(A)(\psi.A\psi)}.$$

In the foregoing discussion we have made three assertions which we now prove.

(1)
$$e_{rs}^{i}e_{uv}^{j} = \frac{n_i n_j}{N^2}\sum_{A,B}ABU_{sr}^{i}(A^{-1})U_{vu}^{j}(B^{-1})\,;$$

now write

$$AB = C, \quad \text{then} \quad B^{-1} = C^{-1}A \quad \text{and} \quad U_{vu}^{j}(B^{-1}) = \sum_{t}U_{vt}^{j}(C^{-1})U_{tu}^{j}(A).$$

With these substitutions,

$$e_{rs}{}^{i}e_{uv}{}^{j} = \sum_{l}\left[\frac{n_i}{N}\sum_{A}U_{sr}{}^{i}(A^{-1})\,U_{lu}{}^{j}(A)\right]\left[\frac{n_j}{N}\sum_{C}CU_{vl}{}^{j}(C^{-1})\right]$$
$$= \sum_{l}\delta_{ij}\,\delta_{rl}\,\delta_{su}\,e_{lv}{}^{j} = \delta_{ij}\,\delta_{su}\,e_{rv}{}^{j},$$

using the generalized orthogonality relations.

(2) Given $\quad (e_{rs}{}^{i}\psi) \neq 0, \quad$ then $\quad (e_{ls}{}^{i}\psi) \neq 0, \quad t = 1 \ldots n_i,$

for $(e_{ts}{}^{i}\psi) = 0$ implies $(e_{rs}{}^{i}\psi) = e_{rt}{}^{i}(e_{ts}{}^{i}\psi) = 0$, contrary to hypothesis.

(3) In order to prove the assertion that given non-zero vectors $(e_{rp}{}^{i}\psi)$ and $(e_{rq}{}^{i}\phi)$, then either $(e_{sq}{}^{i}\phi) = c(e_{sp}{}^{i}\psi)$, $s = 1 \ldots n_i$, or the n_i vectors $e_{sq}{}^{i}\phi$, $s = 1 \ldots n_i$, are linearly independent of the set $\{e_{sp}{}^{i}\psi\}$, we note first that by our definition of irreducibility, two spaces S_p and S_q, irreducible under a group G, are either identical or possess no non-zero vector X in common ; for suppose $X \in S_p$ and $X \in S_q$ then both S_p and S_q are spanned by the same set of vectors $\{AX\}$, $A \in G$; that is $S_p \equiv S_q$. The precise nature of the linear dependence follows from the observation that if the vector $(e_{sq}{}^{i}\phi)$ is contained in the space with basis $\{e_{sp}{}^{i}\psi\}$, then $(e_{sq}{}^{i}\phi) = c(e_{sp}{}^{i}\psi)$, for $(e_{sq}{}^{i}\phi)$ is orthogonal to $(e_{tp}{}^{i}\psi)$ if $t \neq s$, since

$$(e_{sq}{}^{i}\phi).(e_{tp}{}^{i}\psi) = \phi.e_{qs}{}^{i}e_{tp}{}^{i}\psi = (\phi.e_{qp}{}^{i}\psi)\,\delta_{st}.$$

Finally, given

$$(e_{sq}{}^{i}\phi) = c(e_{sp}{}^{i}\psi), \quad \text{then} \quad (e_{tq}{}^{i}\phi) = c(e_{tp}{}^{i}\psi), \quad t = 1 \ldots n_i,$$

for

$$(e_{tq}{}^{i}\phi) = e_{ts}{}^{i}e_{sq}{}^{i}\phi = c\,e_{ts}{}^{i}e_{sp}{}^{i}\psi = c(e_{tp}{}^{i}\psi).$$

PART 2

Introduction

We begin Part 2 with a commentary on the symmetry properties of crystal lattices, the purpose being to present a familiar subject to the reader, as an exercise in the abstract language developed in Part 1. For the most part, we shall not prove our assertions in this commentary, but leave their proof to the imagination of the reader.

It would not be possible to describe adequately in one article all the applications of group theory to solid state physics. Certain closely related basic topics have been selected therefore, which are in most need of detailed exposition. The topics chosen are :

(*a*) the symmetry group of the Dirac Hamiltonian ;

(*b*) Wigner time-reversal symmetry;

(*c*) symmetry in the many-electron problem ;

(*d*) the irreducible representations of space groups containing glide-planes and screw axes.

§1. Lattices and Space Groups

Given a basis $\{a_r\}$ in an N-dimensional space S, we can define the corresponding lattice L as that infinite sub-set of S consisting of all vectors of the form $n = \Sigma_r a_r n_r$, where the coordinates n_r are integers. If we assume a real metric tensor

$$g_{rs} = (a_r . a_s) = (a_s . a_r) = g_{sr},$$

we can define the point group of the corresponding ' empty ' lattice $L \equiv \{\mathbf{n}\}$, as the set G_0 of unitary operators α such that $\mathbf{n} \in L$ implies $\alpha n \in L$. The matrix representation of G_0, referred to the basis $\{a_r\}$, is the set of $N \times N$ matrices under which the tensor $\{g_{rs}\}$ is invariant, and whose elements $\{U_{rs}(\alpha)\}$ are integers—for the invariance condition

$$g_{rs} = \sum_{p,q} g_{pq} U_{pr}(\alpha) U_{qs}(\alpha)$$

implies the unitary condition $(\alpha a_r).(\alpha a_s) = (a_r . a_s)$. The set G_0 is a group, for $\alpha \in G_0$, $\beta \in G_0$ implies $\alpha\beta \in G_0$, while $\alpha \in G_0$ implies $\alpha^{-1} \in G_0$ since the elements of the matrix $U(\alpha^{-1})$ are also integers. It is evident that the existence and structure of these point groups stem from the properties of the integers. Among the set of $N \times N$ matrices over the integers, there exist sub-sets which possess the group property. The problem of discovering these crystal point groups in the abstract is essentially a problem in the theory of numbers.

The basis $\{a_r\}$ of a given lattice $L \equiv \{\mathbf{n}\}$ is not unique if we can find, as we always can, a basis transformation matrix T such that all the matrix elements of T and T^{-1} are integers. We do *not* require that T shall correspond to a unitary operator. It follows from the method of constructing the reciprocal of a matrix that our conditions are satisfied if the elements of T are integers and $\det T = \pm 1$. Since the basis of $L \equiv \{\mathbf{n}\}$ is not unique, neither is the metric tensor, but the point group $G_0 \equiv \{\alpha\}$ is unique.

The point group of the empty lattice always contains J, the operation of inversion in the origin, defined by $Ja_r = -a_r$, for all r. Specializing to three dimensions, if we interpret the operations α geometrically, we find that α is a proper (rigid) rotation of the lattice L into itself about some axis, if the determinant $|U_{rs}(\alpha)| = +1$ while α is some combination of a pure rotation with the inversion J if $\det U(\alpha) = -1$. The unitary condition on α forces $|\det U(\alpha)| = 1$, while the integral condition on the matrix elements $U_{rs}(\alpha)$ makes $\det U(\alpha)$ real, and limits the angular magnitude of the rotations to $\pm\pi/3$, $\pm\pi/2$, $\pm 2\pi/3$ and π radians. (In certain problems, e.g. molecules, local crystal fields, we are not concerned with the translational symmetry of a lattice and other rotations are admissible, for example $2\pi/5$ in the icosahedral group.)

The seven crystal systems with which we are familiar in three dimensions manifest themselves in our formalism when we discover that we can distinguish seven and, only seven, non-equivalent structures of 3×3 real symmetric tensors ; that is, structures which cannot be transformed into one another by a change of

basis in L. The conventional forms for these tensors and the names of the corresponding lattices are

Triclinic	Monoclinic	Orthorhombic	Tetragonal
$g_{11}\ g_{12}\ g_{13}$	$g_{11}\ g_{12}\ 0$	$g_{11}\ 0\ 0$	$a\ 0\ 0$
$g_{21}\ g_{22}\ g_{23}$	$g_{21}\ g_{22}\ 0$	$0\ g_{22}\ 0$	$0\ a\ 0$
$g_{31}\ g_{32}\ g_{33}$	$0\ 0\ g_{33}$	$0\ 0\ g_{33}$	$0\ 0\ c$

Trigonal (rhombohedral)	Hexagonal	Cubic
$a\ b\ b$	$a\ -a/2\ 0$	$a\ 0\ 0$
$b\ a\ b$	$-a/2\ a\ 0$	$0\ a\ 0$
$b\ b\ a$	$0\ 0\ c$	$0\ 0\ a$

In the case of the triclinic, monoclinic, and trigonal tensors we can, by special choice of the parameters g_{rs}, raise the symmetry group of the metric tensor to one corresponding to a lattice of higher symmetry. For example, there are two trigonal lattices with cubic symmetry, corresponding to the two choices, $b = -a/3$ (body-centred cubic), and $b = a/2$ (face-centred cubic). We thus distinguish the complete set of 14 Bravais lattices.

The 32 crystal point groups.

The seven point groups of the empty three-dimensional lattices, and their subgroups, may be derived from the structures of the eleven pure rotation groups (det $\alpha = +1$), denoted by

$$C_1 C_2 C_3 C_4 C_6,\ D_2 D_3 D_4 D_6,\ T \text{ and } 0,$$

in Schonflies' notation (Kittel 1953). Given these eleven groups, C_n, D_n, T and 0, we can construct a further eleven groups by augmenting each of these with the inversion operator J. Since $J^2 = \epsilon$ the unit operator, and $\alpha J = J\alpha$ for all α, each point group so formed is a direct product group of a pure rotation group with the group (ϵ, J). It follows, from our discussion in §6, Part 1, that their irreducible representations are defined by those of the pure rotation groups, together with the parity of the representation with respect to the inversion J (even or odd). The remaining $10 = 32 - 22$ point groups are isomorphous with certain of the pure rotation groups, being obtained from them by replacing particular diad axes (180° rotations) by mirror planes. It is sufficient therefore to publish the character tables for the eleven pure rotation groups (Margenau and Murphy 1943).

Space group operators.

Given a space S containing a lattice $\mathrm{L} \equiv \{\mathbf{n}\}$ with point group G_0, we can define operators $(\alpha|\mathbf{n})$ on arbitrary vectors $X \in \mathrm{S}$, where α is an element of G_0, by the relations $(\alpha|\mathbf{n})X = \alpha X + \mathbf{n}$,

$$\alpha X = \alpha \sum_r a_r X_r = \sum_{r,s} a_s U_{sr}(\alpha) X_r.$$

We write the unit operator in the form $(\epsilon|0)$, while an operator of the form $(\epsilon|\mathbf{n})$ will be called a pure translation operator. The product of two such operators $(\alpha|\mathbf{n})$ and $(\beta|\mathbf{m})$ follows at once from this definition, for

$$(\alpha|\mathbf{n})(\beta|\mathbf{m})X = (\alpha|\mathbf{n})(\beta X+\mathbf{m}) = (\alpha|\mathbf{n})Y \quad \text{where} \quad Y = \beta X+\mathbf{m},$$

$$= \alpha Y+\mathbf{n} = \alpha\beta X+\alpha\mathbf{m}+\mathbf{n} = (\alpha\beta|\mathbf{n}+\alpha\mathbf{m})X.$$

Furthermore, $(\alpha|\mathbf{n})^{-1} = (\alpha^{-1}|-\alpha^{-1}\mathbf{n})$, for

$$(\alpha^{-1}|-\alpha^{-1}\mathbf{n})(\alpha|\mathbf{n}) = (\epsilon|0) = (\alpha|\mathbf{n})(\alpha^{-1}|-\alpha^{-1}\mathbf{n}).$$

It follows from these relations that the set $\{(\alpha|\mathbf{n})\}$ is an infinite discrete group, the space group $\mathscr{S}_0$ of the empty lattice L. The sub-group $\{(\epsilon|\mathbf{n})\}$ of $\mathscr{S}_0$ is called the translation sub-group T. We can construct further sub-groups $\mathscr{S}'$, such that $T\subset\mathscr{S}'\subset\mathscr{S}_0$, by choosing the set $\{\alpha\}$ to be a sub-group of G_0.

The space S, in which the lattice L is embedded, contains infinite sets of planes, lines, and special points, defined by the solutions in X of equations of the form $(\alpha|\mathbf{n})X = X$. These are the symmetry elements that define the space group geometrically.

Space groups of the form $\{(\alpha|\mathbf{n})\}$ are such that the lattice L is invariant under $\{(\alpha|\mathbf{n})\}$, for $(\alpha|\mathbf{n})\mathbf{m} = (\alpha\mathbf{m}+\mathbf{n})\in$ L. We now generalize the structure of our space groups by admitting operations of the form $(\alpha|\mathbf{v}(\alpha)+\mathbf{n})$ where $\mathbf{v}(\alpha)\notin$ L, but satisfies the condition that $\{(\alpha|\mathbf{v}(\alpha)+\mathbf{n})\}$ shall be an infinite discrete group, under the multiplication rule

$$(\alpha|\mathbf{v}(\alpha)+\mathbf{n})(\beta|\mathbf{v}(\beta)+\mathbf{m}) = (\alpha\beta|\mathbf{v}(\alpha)+\mathbf{n}+\alpha\mathbf{v}(\beta)+\alpha\mathbf{m}).$$

This condition implies restrictions on the vectors $\mathbf{v}(\alpha)$, for example,

$$\mathbf{v}(\alpha\beta) = \mathbf{v}(\alpha)+\alpha\mathbf{v}(\beta)+\mathbf{n}'$$

$$\mathbf{v}(\beta\alpha) = \mathbf{v}(\beta)+\beta\mathbf{v}(\alpha)+\mathbf{m}'.$$

A space group of the form $\{(\alpha|\mathbf{v}(\alpha)+\mathbf{n})\}$ may be, however, a simple space group in disguise; for example, if $\mathbf{v}(\alpha) = (\epsilon-\alpha)X_0$, where X_0 is a fixed vector, then $\{(\alpha|\mathbf{v}(\alpha)+\mathbf{n})\}$ is the space group $\{(\alpha|\mathbf{n})\}$ referred to an origin at the position X_0. A plane network of regular hexagons illustrates this case. In the non-trivial case, where $\{(\alpha|\mathbf{v}(\alpha)+\mathbf{n})\}$ cannot be reduced to the form $\{(\alpha|\mathbf{n})\}$ by a change of origin, certain of the operations in the space group correspond geometrically to glide-plane or screw symmetry. The space groups of diamond and graphite are examples of this type of structure.

Substitutional space groups.

Given a space group $\{(\alpha|\mathbf{v}(\alpha)+\mathbf{n})\}$, we define corresponding unitary substitutional operations $[\alpha|\mathbf{v}(\alpha)+\mathbf{n}]$ on a function $\phi(\mathbf{X})$ of the argument $\mathbf{X}$, by the relation

$$\phi(\mathbf{X}) = [\alpha|\mathbf{v}(\alpha)+\mathbf{n}]\psi(\mathbf{X}) = \psi[(\alpha|\mathbf{v}(\alpha)+\mathbf{n})^{-1}\mathbf{X})], \quad \text{for all } \mathbf{X}.$$

With this definition, the multiplication rule for the operators $[\alpha|\mathbf{v}(\alpha)+\mathbf{n}]$ is identical to that for the operators $(\alpha|\mathbf{v}(\alpha)+\mathbf{n})$; that is

$$[\alpha|\mathbf{v}(\alpha)+\mathbf{m}][\beta|\mathbf{v}(\beta)+\mathbf{n}] = [\alpha\beta|\alpha(\mathbf{v}(\beta)+\mathbf{n})+\mathbf{v}(\alpha)+\mathbf{m}],$$

for
$$\begin{aligned}[\alpha|\mathbf{v}(\alpha)+\mathbf{m}][\beta|\mathbf{v}(\beta)+\mathbf{n}]\psi(\mathbf{X}) &= [\alpha|\mathbf{v}(\alpha)+\mathbf{m}]\,\theta(\mathbf{X}) \\ &\quad \text{where } \theta(\mathbf{X}) = \psi[(\beta|\mathbf{v}(\beta)+\mathbf{n})^{-1}\mathbf{X}] \\ &= \theta[(\alpha|\mathbf{v}(\alpha)+\mathbf{m})^{-1}\mathbf{X}] \\ &= \psi[(\beta|\mathbf{v}(\beta)+\mathbf{n})^{-1}(\alpha|\mathbf{v}(\alpha)+\mathbf{m})^{-1}\mathbf{X}] \\ &= \psi[(\alpha\beta|\alpha(\mathbf{v}(\beta)+\mathbf{n})+\mathbf{v}(\alpha)+\mathbf{m})^{-1}\mathbf{X}].\end{aligned}$$

Using this formalism, if a function $V(X)$ is invariant under a space group $\{(\alpha|\mathbf{v}(\alpha)+\mathbf{n})\}$, then $[\alpha|\mathbf{v}(\alpha)+\mathbf{n}]V(\mathbf{X}) = V(\mathbf{X})$ for all $\mathbf{X}$, and for all $[\alpha|\mathbf{v}(\alpha)+\mathbf{n}] \in \{[\alpha|\mathbf{v}(\alpha)+\mathbf{n}]\}$. The unitary condition on the operators $[\alpha|\mathbf{v}(\alpha)+\mathbf{n}]$ requires that in the definition of scalar products $(\psi.\phi)$, the functions $\psi(\mathbf{X})$, $\phi(\mathbf{X})$, and the boundary conditions on the integrals defining scalar products in function space, shall satisfy the condition

$$\int \psi^*(\mathbf{X})\phi(\mathbf{X})d\mathbf{X} = \int \{[\alpha|\mathbf{v}(\alpha)+\mathbf{n}]\psi^*(\mathbf{X})\}\{[\alpha|\mathbf{v}(\alpha)+\mathbf{n}]\phi(\mathbf{X})\}\,d\mathbf{X}.$$

Application to wave mechanics.

The operator calculus developed up to this point is sufficient for quantum mechanical problems without spin. In § 2 we generalize the formalism to take account of spin–orbit coupling. Let us consider briefly at this stage the application of the simple formalism to Schrödinger wave mechanics without spin.

Let $H(\mathbf{X}) = T(\mathbf{X}) + V(\mathbf{X})$ denote a Schrödinger one-particle Hamiltonian operator where $T(\mathbf{X})$ denotes the kinetic energy of the particle and $V(\mathbf{X})$ a potential energy function such that $V(\mathbf{X})$ is invariant under the space group $\{[\alpha\,\mathbf{v}|(\alpha)+\mathbf{n}]\}$, then $[\alpha|\mathbf{v}(\alpha)+\mathbf{n}]H = H[\alpha|\mathbf{v}(\alpha)+\mathbf{n}]$, for the kinetic energy operator is isotropic. Since the operators $[\alpha|\mathbf{v}(\alpha)+\mathbf{n}]$ are unitary, it follows from § 5, Part 1, that the eigenvectors and eigenvalues of this Hamiltonian can be classified according to the irreducible representations of the space group $\{[\alpha|\mathbf{v}(\alpha)+\mathbf{n}]\}$, a subject to which we turn in § 5 of Part 2.

§ 2. The Symmetry Group of the Dirac Hamiltonian

The discovery of the symmetry group of operators for an electron system with spin–orbit coupling presents a fascinating and subtle problem. We begin by considering a one-electron system. If we neglect the electron's spin, the symmetry group of the Hamiltonian for an electron in a potential field $V(\mathbf{r})$ is evidently just the group $\mathscr{G}$ of substitutional operations under which the function $V(\mathbf{r})$ is invariant. When we include spin–orbit coupling, we shall find that this group $\mathscr{G}$ is a homomorphic image of the true symmetry group G. It is, in fact, easier to discover this group G if we work in terms of the Dirac four-component formalism, instead of the familiar two-component spin-formalism. For the benefit of the reader who is not familiar with the four-component formalism, we remark that there is nothing magical about it : the foundations for this formalism were laid in Part 1,

§§ 1, 2 and 3. We begin by describing a certain class of operator and vector space in the language of Part 1.

The operators γ_μ.

Consider a four-dimensional space S. We define four linear operators on S by the ten relations,

$$\gamma_\mu\gamma_\lambda+\gamma_\lambda\gamma_\mu = 2I\delta_{\mu\lambda}, \quad \begin{matrix}\mu = 1,2,3,4,\\ \lambda = 1,2,3,4.\end{matrix} \qquad \ldots\ldots(1)$$

The implications of this definition will become apparent later. We now consider a particular representation of the set $\{\gamma_\mu\}$. In order to simplify the manipulation of 4×4 matrices we introduce a condensed notation, by noting that in a relation of the form $Y_r = \Sigma_s A_{rs} X_s$, we can take the coefficients A_{rs} to be $n\times n$ matrices, if we take X_s and Y_r to be ordered sets of n numbers ; for it is evident that we can carry out the indicated operations of multiplication and addition with the generalized quantities A_{rs} and X_s. In this condensed notation, our 4×4 matrices become 2×2 matrices whose elements are 2×2 sub-matrices, while our vectors become two-dimensional vectors whose coordinates are ordered pairs of numbers. In our case it is sometimes convenient to go one step further, and replace the 2×2 sub-matrices by certain linear operators. To this end, we introduce the three Pauli operators σ_j, $j = 1,2,3$, defined by the Hermitian condition $\sigma_j^\dagger = \sigma_j$, and the relations

$$\left.\begin{aligned}\sigma_j\sigma_k+\sigma_k\sigma_j &= 2i\,\delta_{jk},\\ \sigma_j\sigma_k-\sigma_k\sigma_j &= 2i\,\sigma_l \quad (j\neq k).\end{aligned}\right\} \qquad \ldots\ldots(2)$$

When we require an explicit 2×2 matrix representation of $\{\sigma_j\}$,we shall use the conventional representation

$$\sigma_1 = \begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix} \quad \sigma_2 = \begin{pmatrix}0 & -i\\ i & 0\end{pmatrix} \quad \sigma_3 = \begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}. \qquad \ldots\ldots(3)$$

In terms of these Pauli operators, the 4×4 matrix representation of the set $\{\gamma_\mu\}$ which we require, relative to an orthonormal basis $\{e_n\}$ in the space S, is

$$\|\gamma_k\| = \begin{pmatrix}0 & -i\sigma_k\\ i\sigma_k & 0\end{pmatrix} \quad k = 1,2,3, \quad \|\gamma_4\| = \begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}; \qquad \ldots\ldots(4)$$

where, according to our taste, we can either regard the Pauli operators as symbols for the given 2×2 matrices, or we can boldly work in terms of the σ_j as linear operators, using the defining relations where necessary. In this way we can confirm, using 2×2 matrix algebra and the given representation of $\{\gamma_\mu\}$ in terms of $\{\sigma_k\}$, that $\|\gamma_\mu\gamma_\lambda+\gamma_\lambda\gamma_\mu\| = 2\|I\|\delta_{\mu\lambda}$. We have, for example,

$$\|\gamma_j\gamma_k+\gamma_k\gamma_j\| = \begin{pmatrix}(\sigma_j\sigma_k+\sigma_k\sigma_j) & 0\\ 0 & (\sigma_j\sigma_k+\sigma_k\sigma_j)\end{pmatrix} = 2\|I\|\delta_{jk}.$$

The property of the set $\{\gamma_\mu\}$ with which we shall be most concerned is, given a 4×4 matrix $\|a_{\mu\nu}\|$ with the property $\Sigma_\nu a_{\mu\nu}a_{\lambda\nu} = \delta_{\mu\lambda}$, that there exist corresponding operators T on the space S, defined by the relations

$$\sum_\lambda a_{\mu\lambda}\gamma_\lambda = T^{-1}\gamma_\mu T, \quad \mu = 1,2,3,4. \qquad \ldots\ldots(5)$$

We shall find, in fact, that the operators T are the symmetry operators we seek, if the set of matrices $\{a\}$ afford a representation of the substitutional operations under which the potential energy function of the electron system is invariant. Let us consider, then, the Dirac equation of motion for an electron in an electromagnetic field.

The Dirac equation of motion.

For our purpose it is an advantage to start from the Dirac equation which is symmetrical with respect to the space and time coordinates. There are several conventions in use for the space–time formalism. We shall use the position vector $\mathbf{X}$ with coordinates $(X_1, X_2, X_3, X_4 = ict)$ in the space–time continuum, and the Euclidean metric defined by $(X.X) = X_\mu X_\mu$. The dummy suffix notation will be adopted here and in § 3 on time-reversal symmetry; that is, a repeated suffix implies summation. The Lorentz transformations

$$X_\mu' = a_{\mu\nu} X_\nu, \qquad \ldots\ldots(6)$$

on the coordinates of $\mathbf{X}$, are defined by the conditions

$$\left.\begin{aligned} &a_{\mu\nu} a_{\lambda\nu} = \delta_{\mu\lambda} = a_{\nu\mu} a_{\nu\lambda}, \\ &\left.\begin{matrix} a_{jk} \\ a_{44} \end{matrix}\right\} \text{ real, } \quad j = 1,2,3\,; \quad k = 1,2,3. \\ &\left.\begin{matrix} a_{k4} \\ a_{4k} \end{matrix}\right\} \text{ pure imaginary, } \quad k = 1,2,3. \end{aligned}\right\} \qquad \ldots\ldots(7)$$

With this formalism, the time-like component $A_4(\mathbf{X})$ of the four-potential, representing the electromagnetic field seen by the electron, is of the form $A_4(\mathbf{X}) = i\phi(\mathbf{X})/c$, where $\phi(\mathbf{X})$ is the scalar potential. Since the four-potential satisfies the Lorentz condition

$$\frac{\partial}{\partial X_\nu} A_\nu(\mathbf{X}) = 0, \quad \text{and} \quad \frac{\partial}{\partial X_\nu} = \frac{\partial}{\partial X_\mu'} a_{\mu\nu},$$

we have the relations $A_\nu(\mathbf{X}) = A_\lambda'(\mathbf{X}) a_{\lambda\nu}$, where the functions $A_\lambda'(\mathbf{X})$ are the coordinates of the four-potential referred to the primed coordinate system.

The carrier space of the Dirac equation is a product space constructed by combining a function space with our four-dimensional space S on which we operate with the γ's. As we noted at the beginning of Part 1, the general vector Ψ in our carrier space is then of the form $\Psi = e_n \psi^n$. (The usual spinor notation is inconvenient here, because it suppresses the space S on which we wish to operate explicitly with the symmetry operators T.) With these conventions, the equation of motion for an electron with charge $-e$ takes the form

$$\left[\gamma_\mu\left(\hbar \frac{\partial}{\partial X_\mu} + ieA_\mu(\mathbf{X})\right) + mc\right] \Psi = 0, \qquad \ldots\ldots(8)$$

where the effect of the operator $\gamma_\mu[\hbar(\partial/\partial X_\mu) + ieA_\mu]$ on the vector $\Psi = e_n \psi^n$ is defined to be

$$\gamma_\mu\left(\hbar \frac{\partial}{\partial X_\mu} + ieA_\mu\right) e_n \psi^n = (\gamma_\mu e_n)\left(\hbar \frac{\partial}{\partial X_\mu} + ieA_\mu\right) \psi^n.$$

Given a matrix representation of $\{\gamma_\mu\}$, we can expand the Dirac equation into a set of four simultaneous differential equations in terms of the functions ψ^ν. As a demonstration of the interpretation of the notation, we now transform the Dirac equation into the familiar Pauli–Schrödinger eigenvalue equation for an electron in an electromagnetic field. To do this, we introduce our condensed representation of the vector Ψ and the operators $\{\gamma_\mu\}$. We write $\Psi_1 = (\psi^1, \psi^2)$, $\Psi_2 = (\psi^3, \psi^4)$, and substitute $(\partial/\partial X_4) = -(i/c)(\partial/\partial t)$, $A_4 = (i\phi/c)$. We have then the condensed 2×2 matrix equation :

$$\begin{pmatrix} mc - \frac{1}{c}\left(i\hbar\frac{\partial}{\partial t} + e\phi\right) & \sigma_k(p_k + eA_k) \\ -\sigma_k(p_k + eA_k) & mc + \frac{1}{c}\left(i\hbar\frac{\partial}{\partial t} + e\phi\right) \end{pmatrix} \begin{vmatrix} \Psi_1 \\ \Psi_2 \end{vmatrix} = 0, \qquad \ldots\ldots(9)$$

which represents the pair of equations

$$\left[mc^2 - \left(i\hbar\frac{\partial}{\partial t} + e\phi\right)\right]\Psi_1 + c\sigma_k(p_k + eA_k)\Psi_2 = 0,$$

$$c\sigma_k(p_k + eA_k)\Psi_1 - \left[mc^2 + \left(i\hbar\frac{\partial}{\partial t} + e\phi\right)\right]\Psi_2 = 0,$$

where $p_k = -i\hbar(\partial/\partial X_k)$. Since we require an energy eigenvalue equation we replace $i\hbar(\partial/\partial t)$ by the energy E. We have then

$$\left.\begin{aligned} (E + e\phi - mc^2)\Psi_1 &= c[\sigma_k(p_k + eA_k)]\Psi_2, \\ c[\sigma_k(p_k + eA_k)]\Psi_1 &= (E + e\phi + mc^2)\Psi_2. \end{aligned}\right\} \qquad \ldots\ldots(10)$$

In the non-relativistic limit $E \to mc^2$, and the components ψ^2, ψ^3 become small relative to ψ and ψ^2. Eliminating Ψ_2 between the pair of equations (10), we have the relation

$$(E - mc^2 + e\phi)\Psi_1 = c^2[\sigma_k(p_k + eA_k)](E + mc^2 + e\phi)^{-1}[\sigma_k(p_k + eA_k)]\Psi_1.$$

If we neglect the term $e\phi$ on the right-hand side and make the non-relativistic approximation $E = W + mc^2$, where $W \ll mc^2$, this equation becomes, using the relations (2),

$$\left.\begin{aligned} (W + e\phi)\Psi &= \frac{1}{2m}[\sigma_j(p_j + eA_j)][\sigma_k(p_k + eA_k)]\Psi \\ &= \left[\frac{1}{2m}(\mathbf{p} + e\mathbf{A})^2 + \frac{e\hbar}{2m}\boldsymbol{\sigma}.\operatorname{curl}\mathbf{A}\right]\Psi, \end{aligned}\right\} \qquad \ldots\ldots(11)$$

which is the usual non-relativistic Pauli–Schrödinger eigenvalue equation, without the spin–orbit coupling term, which we neglected when we dropped the term $e\phi$ above.

Symmetry operators.

The Dirac equation is defined to be invariant in form under a Lorentz transformation of coordinates. This is the clue to the discovery of the symmetry

operators we seek. Consider then, the transformation of (8) to a new set of coordinates $\{X_\mu'\}$ defined by (6). Let $\{\psi'^n\}$ denote the set of functions such that

$$\psi'^n(X_1', X_2', X_3', X_4') = \psi^n(X_1, X_2, X_3, X_4), \quad n = 1, 2, 3, 4, \quad \ldots\ldots(12)$$

then in terms of the primed coordinate system, (8) becomes

$$\left[a_{\nu\mu}\gamma_\mu\left(\hbar\frac{\partial}{\partial X_\nu'} + ieA_\nu'(\mathbf{X})\right) + mc\right] e_n\psi'^n(\mathbf{X}) = 0. \quad \ldots\ldots(13)$$

Now corresponding to the Lorentz transformation $\|a_{\nu\mu}\|$, there exist operators T on the space S, such that $a_{\nu\mu}\gamma_\mu = T^{-1}\gamma_\nu T$. (We demonstrate the existence of these operators T by solving these equations for T.) Making this substitution for $a_{\nu\mu}\gamma_\mu$ in (13), and noting $mc = T^{-1}Tmc$, we obtain

$$T^{-1}\left[\gamma_\nu\left(\hbar\frac{\partial}{\partial X_\nu'} + ieA_\nu'(\mathbf{X})\right) + mc\right](Te_n)\psi'^n(\mathbf{X}) = 0 \quad \ldots\ldots(14)$$

which, on multiplication from the left by T, becomes

$$\left[\gamma_\nu\left(\hbar\frac{\partial}{\partial X_\nu'} + ieA_\nu'(\mathbf{X})\right) + mc\right](Te_n)\psi'^n(\mathbf{X}) = 0. \quad \ldots\ldots(15)$$

We can now drop the prime on the coordinates, since both sets of coordinates range over the same domain. We have then

$$\left[\gamma_\nu\left(\hbar\frac{\partial}{\partial X_\nu} + ieA_\nu'\right) + mc\right] T\Psi = 0, \quad \ldots\ldots(16)$$

where we have generalized the definition of the operator T to include the substitutional transformation operation on the functions $\{\psi^n\}$; that is,

$$T\Psi = (Te_n)\psi'^n. \quad \ldots\ldots(17)$$

It follows from (16) that if our Lorentz transformation is such that A_ν' is functionally identical to A_ν, $\nu = 1, 2, 3, 4$, then Ψ and $T\Psi$ are solutions of the same equation. We are not quite home, however, for we require our symmetry operators to commute with the Hamiltonian of the system. Writing our equation of motion in Hamiltonian form, $i\hbar(\partial/\partial t)\Psi = H\Psi$, what we have shown so far is that

$$i\hbar\frac{\partial}{\partial t}\Psi = H\Psi \quad \text{implies} \quad i\hbar\frac{\partial}{\partial t}(T\Psi) = H(T\Psi)$$

if T is such that $A_\nu' \equiv A_\nu$. It follows that $HT = TH$, if $(\partial/\partial t)(T\Psi) = T(\partial/\partial t)\Psi$; that is, if our Lorentz transformation is an operation on the space coordinates alone ($a_{44} = 1$). It follows that the symmetry operators T are such that the three-vector potential $\mathbf{A}$ and the scalar potential ϕ are each invariant under the corresponding transformations induced by T. In order to preserve normalization of the vectors Ψ, we introduce the further constraint that the operators T shall be unitary.

The generalization of our symmetry operators to include translational symmetry is straightforward. Examination of the foregoing equations shows that a translation operation has no effect on the four-dimensional space S, the only effect is a substitutional operation on the functions $\{\psi^n\}$ and the potential A_ν. It follows

that space-group operations $(\alpha|\mathbf{v}(\alpha)+\mathbf{n})$ on the coordinates $\{X_j\}$ induce corresponding operations T, where the definition of T includes the substitutional transformation operation. These space-groups, known as double space-groups (Elliott 1954), will be discussed in § 5 (Part 2).

In addition to these unitary symmetry operators T, there may exist non-linear symmetry operators, in particular the Wigner time-reversal operator, which we discuss in § 3 (Part 2).

The groups $\{T\}$ and $\{a\}$.

Having discovered our unitary symmetry operators, we now consider the structure of the symmetry group of the system. Let $\{T\}$ and $\{a\}$ denote the complete sets of operators T and Lorentz transformations a, such that

$$a_{\mu\nu}\gamma_\nu = T^{-1}\gamma_\mu T.$$

We can see at once that the sets $\{T\}$ and $\{a\}$ are groups; for suppose that a is the image of T and b the image of R, then $\{T\}$ contains TR and RT, and $\{a\}$ contains ab and ba, for

$$(TR)^{-1}\gamma_\mu(TR) = R^{-1}T^{-1}\gamma_\mu TR = a_{\mu\nu}b_{\nu\lambda}\gamma_\lambda. \qquad \ldots\ldots(18)$$

Furthermore, $T\in\{T\}, a\in\{a\}$ implies $T^{-1}\in\{T\}$ and $a^{-1}\in\{a\}$; for

$$a_{\lambda\mu}{}^{-1}\gamma_\mu = a_{\lambda\mu}{}^{-1}(a_{\mu\nu}T\gamma_\nu T^{-1}) = T\gamma_\lambda T^{-1}. \qquad \ldots\ldots(19)$$

The mapping of $\{T\}$ on $\{a\}$ is many–one, but elements of $\{T\}$ that map on the same image in $\{a\}$ are closely related, for let $R \neq T$ denote two elements of $\{T\}$ that map on the same image in $\{a\}$, then

$$T^{-1}\gamma_\mu T = R^{-1}\gamma_\mu R, \quad \mu = 1,2,3,4,$$

that is,

$$(RT^{-1})\gamma_\mu = \gamma_\mu(RT^{-1}). \qquad \ldots\ldots(20)$$

Now, in fact, the space S is irreducible under the set $\{\gamma_\mu\}$, so it follows from Schur's lemma, that

$$R = xT \qquad \ldots\ldots(21)$$

where x is a scalar. We can, if we prefer, derive this result from (20) by using the explicit matrix representation of $\{\gamma_\mu\}$. The matrix algebra required is simple but too lengthy to give here.

For any given electron system, the groups $\{T\}$ and $\{a\}$ are constrained by the condition that the electromagnetic field of the system shall be invariant under the operations corresponding to $\{a\}$.

The unimodular restriction.

We now ask whether it is necessary to consider every member of the infinite set of operators T, which we can find corresponding to each Lorentz transformation in $\{a\}$. There are three points we have to bear in mind if we propose to limit ourselves to a sub-set $\{T'\}$ of the set $\{T\}$.

(1) The sub-set $\{T'\}$ shall possess the group property.

(2) The sub-group $\{T'\}$ shall be such that if a space spanned by vectors $\{\Psi\}$ is irreducible under $\{T\}$, then the space is irreducible under $\{T'\}$. If this condition were not satisfied, we should not fully exploit the symmetry of the given

Hamiltonian; in particular we should be unable to apply our argument on the splitting of a degenerate energy level by a perturbation, which we discussed in § 5, Part 1.

(3) The classification of spaces irreducible under $\{T'\}$ shall be in one-to-one correspondence with the classification under $\{T\}$. For example, although (2) above were satisfied, we might have two irreducible spaces, equivalent under $\{T'\}$, but inequivalent under the larger group $\{T\}$.

It follows at once from equation (21), that if a space is invariant under an operator T corresponding to a given image a in $\{a\}$, then the space is invariant under all the operators in $\{T\}$ that map on a. Thus (2) above is satisfied so long as $\{T'\}$ contains at least one element corresponding to each element in $\{a\}$. In fact, we shall be compelled to retain more than one element in $\{T'\}$ for each $a \in \{a\}$, in order to satisfy the group condition on $\{T'\}$. Bearing these conditions in mind, we introduce the unimodular restriction on the matrix representative of T in the four-dimensional space S.

A unimodular matrix is a matrix whose corresponding determinant has the value unity. It follows from the rule of multiplication for determinants that the matrix product of two unimodular matrices is itself unimodular. This implies that the group property is preserved if in selecting the sub-set $\{T'\}$, we impose the unimodular restriction on the matrix solutions of the equations $a_{\mu\lambda}\gamma_\lambda = T^{-1}\gamma_\mu T$. Since our solutions are 4×4 matrices, this restriction implies $x^4 = 1$ in equation (21), for $\det(xT) = x^4 \det T$; that is, the mapping of $\{T'\}$ on $\{a\}$ is $4 \to 1$. In fact we shall be able to go further, and show that the restriction $a_{44} = 1$ enables us to condense this mapping to $2 \to 1$. At the same time we shall show that the unimodular constraint satisfies the third point above.

We now consider the problem of calculating the solutions T corresponding to a given Lorentz transformation for which $a_{44} = 1$, $a_{4j} = 0 = a_{j4}$, $j = 1, 2, 3$.

Solutions of the equations $a_{\mu\nu}\gamma_\nu = T^{-1}\gamma_\mu T$, $a_{44} = 1$.

Since $a_{44} = 1, a_{j4} = 0 = a_{4j}$, we have, in particular, $\gamma_4 = T^{-1}\gamma_4 T$ or $T\gamma_4 = \gamma_4 T$; that is, in condensed notation using our explicit representation of T,

$$\begin{pmatrix} P & S \\ R & Q \end{pmatrix} \times \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \times \begin{pmatrix} P & S \\ R & Q \end{pmatrix}.$$

Therefore $R = 0 = S$ and P and Q are unitary 2×2 matrices. This result shows that we may extend the unimodular restriction to each of the matrices P and Q without destroying the group properties of our set $\{T'\}$. Referring back to equation (21), this further restriction implies $x = \pm 1$; that is, each element of $\{a\}$ is an image of two elements in $\{T'\}$. We shall postpone introducing this further restriction until a later stage, where we shall be able to see the implications more clearly.

The remaining three equations $a_{jk}\gamma_k = T^{-1}\gamma_j T$, or $a_{jk}T\gamma_k = \gamma_j T$, now give the six 2×2 matrix equations

$$\left.\begin{aligned} a_{jk}\sigma_k &= P^{-1}\sigma_j Q \\ a_{jk}\sigma_k &= Q^{-1}\sigma_j P \end{aligned}\right\} \quad k = 1, 2, 3. \qquad \ldots\ldots(22)$$

Taken together, these equations imply

$$\sigma_j = (PQ^{-1})\,\sigma_j (PQ^{-1}), \quad j = 1,2,3,$$

which then imply, after several lines of algebra using the explicit representation of $\{\sigma_j\}$,

$$P = \pm Q. \qquad \ldots\ldots(23)$$

The two classes of solution implied by (23) are mutually exclusive, in the sense that the solutions corresponding to a given matrix $\|a_{jk}\|$ are either all of the form $Q = P$ or all of the form $Q = -P$; for suppose a solution of each kind exists for a given $\|a_{jk}\|$, that is

$$-R^{-1}\sigma_j R = a_{jk}\sigma_k = P^{-1}\sigma_j P,$$

then $PR^{-1}\sigma_j + \sigma_j PR^{-1} = 0$, $j = 1,2,3$; but the only solution to the equations (using the representation of $\{\sigma_j\}$)

$$A\sigma_j + \sigma_j A = 0, \quad j = 1,2,3 \qquad \ldots\ldots(24)$$

is

$$A \equiv 0.$$

Therefore one or other of the assumed solutions does not exist. On the other hand, given a solution for $\|a_{jk}\|$, we can at once construct a solution for $\|-a_{jk}\|$ by changing the sign of Q. It follows that it is sufficient to calculate the solutions corresponding to proper rotations, that is $\det a_{jk} = +1$. This we now do. In order to find out which of the two types of equation $a_{jk}\sigma_k = \pm P^{-1}\sigma_j P$ we should solve in this case, we note from equation (24) that no solutions exist to the equations $a_{jk}\sigma_k = -P^{-1}\sigma_j P$, if $a_{jk} = \delta_{jk}$. Since the elements of $\|a_{jk}\|$ are continuous functions of the angle of rotation, we conclude that the solutions $P = Q$ correepond to proper rotations, while the solutions $P = -Q$ correspond to the improper rotations ($\det a_{jk} = -1$), that is

$$a_{jk}\sigma_k = \pm P^{-1}\sigma_j P, \quad \det a_{jk} = \pm 1. \qquad \ldots\ldots(25)$$

Taking the unitary 2×2 matrix P in the general form

$$P \equiv \begin{pmatrix} a & b \\ -b^*\,e^{i\theta} & a^*\,e^{i\theta} \end{pmatrix} \quad \begin{matrix} |a|^2 + |b|^2 = 1, \\ \det P = e^{i\theta}, \end{matrix} \qquad \ldots\ldots(26)$$

where we have still held our hand over the unimodular restriction, the solutions to the equations $a_{jk}\sigma_k = P^{-1}\sigma_j P$ are found to be given by the relations (after some lines of algebra)

$$\left.\begin{aligned} a^2 &= e^{i\theta}\left[\tfrac{1}{2}(a_{11}+a_{22}) + i\tfrac{1}{2}(a_{12}-a_{21})\right], \\ b^2 &= e^{i\theta}\left[\tfrac{1}{2}(a_{22}-a_{11}) + i\tfrac{1}{2}(a_{12}+a_{21})\right], \\ ab &= -e^{i\theta}\tfrac{1}{2}(a_{13} - i\,a_{23}), \quad |a|^2 - |b|^2 = a_{33}. \end{aligned}\right\} \qquad \ldots\ldots(27)$$

If $\|a_{jk}\|$ represents a clockwise rotation through an angle ϕ about an axis with direction-cosines l, m, n, these relations give

$$\left.\begin{aligned} a &= \pm(\cos\tfrac{1}{2}\phi - i\,n\sin\tfrac{1}{2}\phi)\,e^{i\theta/2}, \\ b &= \mp(m + il)\sin\tfrac{1}{2}\phi\,e^{i\theta/2}, \quad |a|^2 + |b|^2 = 1. \end{aligned}\right\} \qquad \ldots\ldots(28)$$

If we impose the unimodular restriction on the matrix P, then $e^{i\theta} = 1$, and (26) gives the expected $2 \to 1$ mapping of $\{T'\}$ on $\{a\}$. In this case, if $\{a\}$ is a crystal point group, the corresponding group $\{T'\}$ is called the corresponding double point group. If we impose the unimodular restriction on T, we have $e^{i\theta} = \pm 1$ and $e^{i\theta/2} = \pm i$. We then have from (26) the expected $4 \to 1$ mapping of $\{T'\}$ on $\{a\}$. There is, however, no advantage in using the larger group $\{T\}$ for the following reason. Let $\{T\}_2$ denote the symmetry group corresponding to the choice $e^{i\theta} = 1$, and let $\{T\}_4$ denote the symmetry group corresponding to $e^{i\theta} = \pm 1$, then $\{T\}_2$ is a sub-group of $\{T\}_4$. The only situation in which there could be an advantage in using $\{T\}_4$ in place of $\{T\}_2$ would be one in which we were faced with two irreducible spaces, equivalent with respect to $\{T\}_2$, but inequivalent with respect to $\{T\}_4$. This situation cannot occur in our context, for in order to define the group $\{T\}_4$ we have defined at the outset, through equation (21), the relations between the elements of $\{T\}_4$ that map on the same image in $\{a\}$. The only irreducible representations of $\{T\}_4$ accessible to us in spaces spanned by our vectors $\Psi = e_n \psi^n$ are those consistent with these defining relations. Given the character with respect to $\{T\}_2$ of a space spanned by vector Ψ, we can calculate from (21) the character with respect to $\{T\}_4$. In our context, in other words, if two spaces are equivalent under $\{T\}_2$, then they are equivalent under $\{T\}_4$. We now observe that the argument is not peculiar to $\{T\}_4$. If two spaces, spanned by eigenvectors of the Dirac Hamiltonian, are equivalent under $\{T\}_2$, then they are equivalent under any larger group $\{T\} \supset \{T\}_2$, if $\{T\}$ and $\{T\}_2$ map on the same image $\{a\}$.

Transformation of spin vectors.

In the Pauli–Schrödinger two-component formalism for an electron, if we neglect spin–orbit coupling, the wave function is the product of a scalar wave function with a two-dimensional spin vector. We note from the derivation of equation (11) that, in the non-relativistic limit, the Pauli–Schrödinger wave function is to be identified with

$$\Psi_1 \equiv (\psi^1, \psi^2) \equiv \psi |a\rangle$$

where $|a\rangle$ denotes the spin part of the wave function. It follows from this identification and equation (9) that a rotation corresponding to the coordinate transformation matrix $\|a_{jk}\|$ induces a transformation P on the vector $|a\rangle$, where P satisfies equations (25), namely $a_{jk}\sigma_k = \pm P^{-1}\sigma_j P$, $\det a_{jk} = \pm 1$. We now show that under this transformation the matrix elements $\langle a|\sigma_j|b\rangle$, $j = 1, 2, 3$, of the Pauli operators $\{\sigma_j\}$, taken between arbitrary spin vectors $|a\rangle$ and $|b\rangle$, transform like the coordinates of an axial vector. Let $|a'\rangle = P|a\rangle$ and $|b'\rangle = P|b\rangle$. We have then at once, since P is unitary

$$\begin{aligned}\langle a'|\sigma_j|b'\rangle &= \langle a|P^{-1}\sigma_j P|b\rangle \\ &= \pm a_{jk}\langle a|\sigma_k|b\rangle \text{ according as } \det a_{jk} = \pm 1 ; \qquad \ldots\ldots (29)\end{aligned}$$

that is, the components of the vector $\langle a|\boldsymbol{\sigma}|b\rangle$ transform like the components of an axial vector. The vector $\langle a|\boldsymbol{\sigma}|b\rangle$, where $|a\rangle \neq |b\rangle$, does not possess a direction in the geometrical sense, because the coordinates are complex numbers; but if

$|a\rangle = |b\rangle$, the vector $\langle a|\sigma|a\rangle$ does possess a direction, determined by the direction cosines

$$\left.\begin{aligned} l &= a_1 a_2^* + a_1^* a_2 = \langle a|\sigma_1|a\rangle, \\ m &= i(a_1 a_2^* - a_1^* a_2) = \langle a|\sigma_2|a\rangle, \\ n &= |a_1|^2 - |a_2|^2 = \langle a|\sigma_3|a\rangle, \end{aligned}\right\} \quad \ldots\ldots(30)$$

where a_1 and a_2 are the complex coordinates of the normalized spin vector $|a\rangle$; that is, $|a_1|^2 + |a_2|^2 = 1$.

The groups U_2 and R_3.

The group U_2 is the group of all unitary unimodular 2×2 matrices over the complex numbers. The group R_3 is the group of all real unitary 3×3 matrices, with unit determinant. These two groups are related through our equations $a_{jk}\sigma_k = P^{-1}\sigma_j P$, if we impose the unimodular constraint on P. Given $P \in U_2$, we can calculate the matrix $\|a_{jk}\|$ in R_3, while given $\|a_{jk}\|$ in R_3, we can calculate the two corresponding members of U_2. Since we can show, by means of the Euler factorization, that every real unitary 3×3 matrix corresponds to some rigid rotation of our three-dimensional space, it follows that R_3, as defined above, is isomorphic with the group of all such rigid rotations.

We mention here as a matter of general interest that every matrix U in U_2 is related to a diagonal member

$$D(\theta) = \begin{pmatrix} e^{-i\theta/2} & 0 \\ 0 & e^{+i\theta/2} \end{pmatrix}, \quad 0 < \theta < 4\pi,$$

of U_2 by a relation of the form

$$U = V^{-1} D(\theta) V,$$

where V is some member of U_2. It follows that the classes of U_2 can be labelled with the parameter $0 < \theta < 4\pi$. It is conventional in quantum mechanics to distinguish the irreducible representation of U_2 by the symbol $\mathscr{D}_j$ where $j = 0, \frac{1}{2}, 1, \frac{3}{2} \ldots$. The spur of the class θ in the irreducible representation $\mathscr{D}_j$ of U_2 is

$$\chi^j(\theta) = \frac{\sin(j+\frac{1}{2})\theta}{\sin\frac{1}{2}\theta}.$$

Since R_3 is a homomorphic image of U_2, we know that the irreducible representations of R_3 are irreducible representations of U_2. These representations are in fact the representations $\mathscr{D}_j$ corresponding to integral values of the parameter j. For integral j we have at once $\chi(\theta) = \chi(\theta + 2\pi)$.

Perturbation of spherical symmetry.

If an electron system has spherical symmetry, the states are classified according to the irreducible representations of U_2. If the number of electrons is odd, only those representations $\mathscr{D}_j$ corresponding to half-integral values of j occur, while if the number of electrons is even, the only irreducible representations $\mathscr{D}_j$ that occur correspond to integral values of j. If an otherwise spherically symmetric system is perturbed by a potential field with the symmetry of one of the crystal point groups $\{a\}$, the perturbed states will belong to irreducible representations of the corresponding group $\{T\}_2$ which is a sub-group of U_2. Using the irreducible characters of

the group $\{T\}_2$ of the system, and the formula for $\chi^j(\theta)$ given above, we can calculate the splitting of the degenerate energy levels under the given perturbation, using the arguments of Part 1, § 6, and the theory of time-reversal degeneracy discussed in Part 2, § 3.

The spin–rotation group.

The group U_2 is sometimes called the spin–rotation group, for as we have seen in (29), the matrices of U_2 are the representatives of the operators P on the spin vectors $|a\rangle$, induced by rotations $\|a_{jk}\|$ of the orientation of the physical system.

Representations.

We now consider a peculiar relation between the irreducible representations of $\{T\}_2$ and of $\{a\}$. Since $\{a\}$ is a homomorphic image of $\{T\}_2$, the irreducible representation of $\{a\}$ will appear among the complete set of those of $\{T\}_2$. It follows, however, from equation (21), that these common irreducible representations cannot appear in a space spanned by our vectors Ψ; for (21) requires that the representatives of two elements R and T, $R = -T$, in $\{T\}_2$ that map on the same image in $\{a\}$, shall differ in sign, which is clearly impossible if the representation of $\{T\}_2$ is simultaneously a representation of $\{a\}$. This is a one-electron argument. When we consider symmetry operations for a many-electron system, we shall find that the argument above is relevant if the number of electrons in the system is odd, but if the number of electrons is even, then the only irreducible representations of $\{T\}_2$ accessible to us are those common to $\{T\}$ and $\{a\}$.

The next point to note is that if we work in terms of the Pauli two-component formalism, that is, in terms of $\Psi_1 \equiv (\psi_1, \psi_2)$, then our symmetry group is $\{P\}$, where P is determined by the relations $a_{jk}\sigma_k = \pm P^{-1}\sigma_j P$, $\det a_{jk} = \pm 1$. We can see from the structure of these relations that there is no distinction in $\{P\}$, between proper ($\det a_{jk} = +1$) and improper ($\det a_{jk} = -1$) rotations. We obtain the same solutions in P, irrespective of whether a is a proper or improper rotation. It follows that within the Pauli formalism it is sufficient to consider the double point groups which map on the eleven pure rotation crystal point groups. A faithful 2×2 matrix representation of these eleven 'proper' double point groups can be obtained using equations (28). The irreducible characters of nearly all these groups were given by Bethe (1929), the remainder by Opechowski (1940). The generalization to double space groups is clearly straightforward, all we have to do is to replace the geometrical point group $\{a\}$ by the induced group $\{P\}$.

If we work in terms of the Dirac four-component formalism we find that there is a distinction between proper and improper rotations, since the transformation on ψ^3 and ψ^4 is opposite in sign to the transformation on ψ^1 and ψ^2, if $\det a_{jk} = -1$.

Symmetry operator for a many-electron system.

We do not know the generalization of Dirac's equation to a many-electron system, so we define a generalization of our symmetry operators T and P, and justify our definition by comparison between prediction and experiment. The state space of our model of a many-electron system is a product of the one-electron spaces, these one-electron spaces being spanned either by two- or four-component spinors, depending on which formalism we choose. The electron–electron

interaction terms in the non-relativistic Schrödinger–Hamiltonian compel us to carry out the same spatial operations simultaneously in all the one-electron orbital spaces, so we define a many-electron symmetry operation to be the operation P carried out simultaneously on all the one-electron spaces which make up the state space of the system.

We now observe, if the number of electrons in our system is even, that the distinction between the representations of P and $P' = -P$ vanishes, for we have to multiply together an even number of negative signs. It follows that for an even number of electrons the only irreducible representations of $\{P\}$ that can appear are those irreducible representations of $\{P\}$ which are simultaneously irreducible representations of $\{a\}$. It appears as if the symmetry group of the system is $\{a\}$ and not $\{P\}$. If on the other hand we are concerned with an odd number of electrons, then the same argument applies as for one electron, and only those irreducible representations of $\{P\}$ can appear which are not irreducible representations of $\{a\}$.

§ 3. Wigner Time-reversal Symmetry

In Part 1, § 5, we noted that the postulate on degeneracy must be qualified if the Hamiltonian of a system is such that there exist non-linear symmetry operators that may carry eigenvectors in one irreducible eigenvector sub-space into another. We now demonstrate that such an operator exists in the case of an electron system, if there is no external magnetic field present. We do not concern ourselves with the general theory of time-reversal (Bell 1955), but limit ourselves to stationary states.

The Wigner time-reversal operator.

Given that the four-potential $\{A_\mu\}$ in the Dirac equation of motion (8) is time-independent, we can write the eigenvector $\Psi = e_n \psi^n$, belonging to the energy eigenvalue E, in the form

$$\Psi = e_n \phi^n \exp(-iEt/h). \qquad \ldots\ldots(31)$$

With this substitution the equation of motion becomes the time-independent eigenvalue equation

$$\left[\gamma_4\left(-\frac{E}{c}+ieA_4\right)+\gamma_k\left(\hbar\frac{\partial}{\partial X_k}+ieA_k\right)+mc\right]e_n\phi^n = 0. \qquad \ldots\ldots(32)$$

We begin our search for the Wigner time-reversal operator by taking the complex conjugate of this equation. This operation is well defined, if we replace the operators γ_μ by their matrix representatives, and we define $e_n{}^* = e_n$. We have then

$$\left[\gamma_4{}^*\left(-\frac{E}{c}+ieA_4\right)+\gamma_k{}^*\left(\hbar\frac{\partial}{\partial X_k}-ieA_k\right)+mc\right]e_n\phi^{*n} = 0, \qquad \ldots\ldots(33)$$

since A_k, $k = 1, 2, 3$, is real, while $A_4 = i\phi/c$ is pure imaginary. Now there exists a 4×4 matrix R such that for our chosen matrix representation of $\{\gamma_\mu\}$,

$$R^{-1}\gamma_\mu R = \gamma_\mu{}^*; \qquad \ldots\ldots(34)$$

for consider, in particular, the condensed 4×4 matrix

$$R=\begin{pmatrix} S & 0\\ 0 & S\end{pmatrix},\quad \text{where } S \text{ is the } 2\times 2 \text{ matrix } \begin{pmatrix} 0 & 1\\ -1 & 0\end{pmatrix}=i\begin{pmatrix} 0 & -i\\ i & 0\end{pmatrix}=i\sigma_2. \qquad \ldots\ldots(35)$$

We have then for our chosen representation of $\{\gamma_\mu\}$,

$$R^{-1}\gamma_k R=\begin{pmatrix} -i\sigma_2 & 0\\ 0 & -i\sigma_2\end{pmatrix}\begin{pmatrix} 0 & -i\sigma_k\\ i\sigma_k & 0\end{pmatrix}\begin{pmatrix} i\sigma_2 & 0\\ 0 & i\sigma_2\end{pmatrix}=\begin{pmatrix} 0 & -i\sigma_2\sigma_k\sigma_2\\ i\sigma_2\sigma_k\sigma_2 & 0\end{pmatrix}=\gamma_k{}^* \qquad \ldots\ldots(36)$$

using the relations $\sigma_j\sigma_k+\sigma_k\sigma_j=2I\delta_{jk}$. Furthermore

$$R^{-1}\gamma_4 R=\gamma_4=\gamma_4{}^* \quad \text{since} \quad \|\gamma_4\| \text{ is real.}$$

It is evident from our discussion of the equation $a_{\mu\nu}\gamma_\nu=T^{-1}\gamma_\mu T$, that the matrix R defined by the equations (34) is not unique, but all the solutions are scalar multiples of one another. Substituting $\gamma_\mu{}^*=R^{-1}\gamma_\mu R$ into (33) and multiplying from the left with R, we obtain the result

$$\left[\gamma_4\left(-\frac{E}{c}+ieA_4\right)+\gamma_k\left(\hbar\frac{\partial}{\partial X_k}-ieA_k\right)+mc\right](Re_n)\phi^{*n}=0. \qquad \ldots\ldots(37)$$

Comparing now (32) and (37), we see, if the magnetic field is zero (that is $A_k\equiv 0$), that $(Re_n)\phi^{*n}$ is a solution of (32) degenerate in energy with the solution $e_n\phi^n$. If we define the operator $\mathscr{R}$ such that

$$\mathscr{R}\Psi=(Re_n)\phi^{*n}=e_m R_{mn}\phi^{*n}, \qquad \ldots\ldots(38)$$

we can assert that $$\mathscr{R}H=H\mathscr{R}, \quad \text{if} \quad A_k\equiv 0. \qquad \ldots\ldots(39)$$

The fact that this symmetry operation involves the operation of taking the complex conjugate in function space means that $\mathscr{R}$ is not unitary and cannot be treated on the same footing as the unitary symmetry operators that form the symmetry group of the Hamiltonian.

Since we do not know the generalization of the Dirac equation to a many-electron system, we assume that the Wigner operation for a many-electron system is simply the operation defined above, applied to each one-electron space simultaneously. Furthermore, it follows from our discussion of the condensation of the Dirac four-component formalism into the Pauli–Schrödinger two-component formalism that the Wigner time-reversal operator in the Pauli formalism is simply $i\sigma_2$, together with the operation of taking the complex conjugate in function space.

We now prove that $\mathscr{R}$ commutes with the elements of the symmetry group $\{T\}_2$. For simplicity, we work in terms of the Pauli two-component formalism, but the generalization to four-component formalism is immediate. We show that $\mathscr{R}$ commutes with $\{P\}$, where we have chosen $e^{i\theta}=1$ in the matrix of equation (26). We note first that the operation of taking the complex conjugate of a function ψ commutes with a substitutional operation on ψ.

Then
$$P\mathscr{R}\Psi=P\mathscr{R}e_n\psi^n=P(e_m R_{mn}\psi^{*n})$$
$$=e_l P_{lm}R_{mn}P\psi^{*n}=e_l P_{lm}R_{mn}(P\psi^n)^*,$$

and
$$\mathscr{R}P\Psi=\mathscr{R}P(e_n\psi^n)=\mathscr{R}[e_m P_{mn}(P\psi^n)]=e_l R_{lm}P_{mn}{}^*(P\psi^n)^*\ ;$$

but $$P_{lm} R_{mn} = R_{lm} P_{mn}^*,$$

for $$\begin{pmatrix} a & b \\ -b^* & a^* \end{pmatrix}\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} -b & a \\ -a^* & -b^* \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} a^* & b^* \\ -b & a \end{pmatrix}.$$

Therefore $$P\mathscr{R} \equiv \mathscr{R}P, \quad P \in \{P\}. \qquad \ldots\ldots(40)$$

A little thought reveals that we can extend this result to the state space of a many-electron system.

We require three more properties of the operator $\mathscr{R}$. We observe that

$$\Psi . \mathscr{R}\Phi = (e_n \psi^n) . (e_l R_{lm} \phi^{*m}) = (\psi^n . \phi^{*m}) R_{nm},$$

while $$\Phi . \mathscr{R}\Psi = (\phi^n . \psi^{*m}) \mathscr{R}_{nm} = (\psi^m . \phi^{*n}) R_{nm} = -(\psi^n . \phi^{*m}) R_{nm},$$

since $R_{nm} = -R_{mn}$ by definition from (35), and $(\phi . \psi^*) = (\psi . \phi^*)$.

Therefore $$\Psi . \mathscr{R}\Phi = -\Phi . \mathscr{R}\Psi.$$

Generalizing to the state space of an $\mathscr{N}$-electron system,

$$\Psi . \mathscr{R}\Phi = (-1)^{\mathscr{N}} \Phi . \mathscr{R}\Psi. \qquad \ldots\ldots(41)$$

The second property required is

$$\mathscr{R}\Psi . \mathscr{R}\Phi = R_{lm} R_{ln} (\psi^{*m} . \phi^{*n}) = \Phi . \Psi, \qquad \ldots\ldots(42)$$

since $$R_{lm} R_{ln} = \delta_{mn}, \quad \text{and} \quad (\psi^{*m} . \phi^{*m}) = \Phi . \Psi.$$

This result can be generalized at once to a many-electron system by examining the definition of a scalar product in Part 1, § 2.

The third property we require is that

$$\mathscr{R}\mathscr{R} = (-1)^{\mathscr{N}}. \qquad \ldots\ldots(43)$$

This follows at once from the definition, for

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}.$$

We observe here that relations (41), (42) and (43), for $\mathscr{N}$ even, are identical to the corresponding relations for the case where we neglect spin–orbit coupling completely, and the time-reversal operator is simply the operation of taking the complex conjugate of the wave function. For this reason, the discussion that follows covers this case as well.

Time-reversal degeneracy.

Let $S^i(E)$ denote an irreducible energy eigenvector sub-space, belonging to the eigenvalue E of the Hamiltonian H of our electron system. Let Ψ denote an eigenvector in $S(E)$, then the vector $\mathscr{R}\Psi$ is itself an eigenvector belonging to the same eigenvalue E, for $\mathscr{R}$ commutes with H (equation (39)). It follows that extra degeneracy is induced by $\mathscr{R}$, if the vector $\mathscr{R}\Psi$ is not contained in the space $S^i(E)$. It follows at once from (41), if $\mathscr{N}$, the number of electrons in the system, is odd and $S^i(E)$ is one-dimensional, that $\mathscr{R}\Psi$ is not contained in $S^i(E)$, for (41) implies that $\mathscr{R}\Psi$ and Ψ are orthogonal. We now consider methods for detecting time-reversal degeneracy.

The first step is to show that the space $\mathscr{R}S^i$ spanned by the set of vectors $\{\mathscr{R}\Psi\}$, where Ψ is an arbitrary vector in $S^i(E)$, is, in fact, an irreducible eigenvector sub-space $S^j(E)$, where the character of $S^j(E)$ with respect to the symmetry group $\{P\}$ of the Hamiltonian is the complex conjugate of the character of $S^i(E)$; that is $\chi^j = (\chi^i)^*$.

Let $\{\Psi_r\}$ denote an orthonormal basis in S^i, where $P\Psi_r = \Psi_s U_{sr}{}^i(P)$, then $\{\mathscr{R}\Psi_r\}$ is an orthonormal basis in the space $\mathscr{R}S^i$; for $\mathscr{R}\Psi_r . \mathscr{R}\Psi_s = \Psi_s . \Psi_r = \delta_{rs}$, from (42). Now from (40)

$$P(\mathscr{R}\Psi_r) = \mathscr{R}(P\Psi_r) = (\mathscr{R}\Psi_s)\, U^*_{sr}{}^i(P).$$

It follows that the character of the space $\mathscr{R}S^i$ is the complex conjugate of the character of S^i. Assuming that $\{P\}$ is a finite group, the irreducibility of the character χ^j of $\mathscr{R}S^i \equiv S^j$ follows from the relation

$$\sum_P |\chi^j(P)|^2 = \sum_P |\chi^i(P)|^2 = \text{ order of } \{P\},$$

since S^i is irreducible. Now S^i and S^j, being irreducible under $\{P\}$, are either identical, or possess no non-zero vector in common; for if X is a non-zero vector common to S^i and S^j, the set of vectors $\{PX\}, P \in \{P\}$, spans both S^i and S^j. It follows at once that time-reversal degeneracy certainly occurs if $(\chi^i)^* \neq \chi^i$. For example, in the case of the cyclic group of order 4, $G \equiv (A, A^2, A^3, A^4 = I)$, which has four irreducible representations, all one-dimensional, with characters

	I	A	A^2	A^3
χ^1	1	1	1	1
χ^2	1	-1	1	-1
χ^3	1	i	-1	$-i$
χ^4	1	$-i$	-1	i

we have degeneracy in pairs between eigenvalues labelled with χ^3 and χ^4. The non-trivial case is therefore that in which χ^i is real, but $\mathscr{R}S^i \neq S^i$. Wigner (1932) showed that for finite groups this case can be detected by calculating the sum $\Sigma_P \chi(P^2)$. Wigner's theorem states, given χ^i real, that if $\mathscr{N}$, the number of electrons in the system, is

$$\left.\begin{array}{ll} (a) \text{ odd, then} & \sum_P \chi^i(P^2) = N \\ (b) \text{ even, then} & \sum_P \chi^i(P^2) = -N \end{array}\right\} \text{ implies } \mathscr{R}S^i \neq S^i, \quad \ldots\ldots(44)$$

where N is the order of the finite group $\{P\}$. In addition, Wigner shows that the case $(\chi^i)^* \neq \chi^i$ can be recognized by the result $\Sigma_P \chi^i(P^2) = 0$. This fact is useful in the case of crystal space groups, to which we shall apply Wigner's theorem in § 5 (Part 2).

We shall see when we derive these theorems, that the dependence on the sign of $(-1)^{\mathscr{N}}$ is a consequence of equation (41). Wigner's theorem follows, once we have proved the relations,

(i) if χ^i is real, then $\Sigma_P \chi^i(P^2) = \pm N$, ……(45)

(ii) $\mathscr{R}S^i = S^i$ implies $\Sigma_P \chi^i(P^2) = N(-1)^{\mathscr{N}}$; ……(46)

for given $\mathscr{N}$ even (odd) and $\Sigma_P \chi^i(P^2) = -N(+N)$, then $\mathscr{R}S^i \neq S^i$, since $\mathscr{R}S^i = S^i$ implies $\Sigma_P \chi^i(P^2) = +N(-N)$, contrary to hypothesis.

It is important to note that we do not assert $\Sigma_P \chi^i(P^2) = +N(-N)$, $\mathscr{N}$ even (odd), implies $\mathscr{R}S^i = S^i$, for this is not true.

Proof of Wigner's formulae.

In order to prove relations (45) and (46), we note that if the character of a representation $\{U^i(P)\}$ is real, then $\{U^i(P)\}$ and the complex conjugate representation $\{U^{*i}(P)\}$, are equivalent representations; that is, there exists a similarity transformation T such that

$$U^{*i}(P) = T^{-1} U^i(P)\, T, \quad P \in \{P\}. \qquad \ldots\ldots(47)$$

This transformation T is not unique, for any scalar multiple of T satisfies (47). We now show that we can choose T to be unitary, that is $T^{-1} = T^\dagger$, if $\{U^i(P)\}$ is an irreducible unitary representation of $\{P\}$. (A proof is necessary, in that two equivalent reducible unitary representations may be related by only non-unitary similarity transformations.) Let $\{U(P)\}$ and $\{V(P)\}$ denote equivalent irreducible unitary representations of $\{P\}$, then

$$V(P)\, V^\dagger(P) = U(I), \quad V(P) = T^{-1} U(P) T,$$

$$U^\dagger(P)\, U(P) = U(I), \quad V^\dagger(P) = T^\dagger U^\dagger(P)(T^{-1})^\dagger = T^\dagger U^\dagger(P)(T^\dagger)^{-1}.$$

It follows that

$$U(I) = V(P)\, V^\dagger(P) = T^{-1} U(P)\, TT^\dagger\, U^\dagger(P)(T^\dagger)^{-1};$$

that is

$$(TT^\dagger)\, U(P) = U(P)(TT^\dagger), \quad P \in \{P\}.$$

Therefore, from the matrix form of Schur's lemma,

$$(TT^\dagger) = cU(I),$$

and since we can multiply T by an arbitrary scalar, we can arrange that

$$TT^\dagger = U(I), \quad \text{that is } T^\dagger = T^{-1}. \qquad \ldots\ldots(48)$$

We now show that (47) and (48) together imply that the matrix T is either symmetric or antisymmetric, that is $T = \pm \hat{T}$; for, choosing $T^{-1} = T^\dagger = \hat{T}^*$ in (47) and rearranging,

$$U^i(P) = TU^{*i}(P)\, \hat{T}^*. \qquad \ldots\ldots(49)$$

Taking the complex conjugate of (49),

$$U^{*i}(P) = T^*\, U^i(P)\, \hat{T}. \qquad \ldots\ldots(50)$$

Combining (49) and (50),

$$U^i(P) = TT^*\, U^i(P)\, \hat{T}\hat{T}^* = TT^*\, U^i(P)(TT^*)^{-1},$$

therefore

$$U^i(P)(TT^*) = (TT^*)\, U^i(P), \quad P \in \{P\}. \qquad \ldots\ldots(51)$$

It follows therefore, from Schur's lemma, that

$$TT^* = cU(I). \qquad \ldots\ldots(52)$$

Now taking the complex conjugate of (48), we have $T^* \tilde{T} = U(I)$, and so, on multiplying (52) on the right with $\tilde{T}$,

$$T = c\tilde{T}. \qquad \ldots\ldots(53)$$

Transposing (53), $$\tilde{T} = cT. \qquad \ldots\ldots(54)$$

Combining (53) and (54), $T = c^2 T$; therefore $c = \pm 1$, and

$$T = \pm \tilde{T}. \qquad \ldots\ldots(55)$$

In order to prove that $\Sigma_P \chi^i(P^2) = \pm N$, we note that

$$\chi^i(P^2) = U_{rs}{}^i(P)\, U_{sr}{}^i(P) = U_{rs}{}^i(P)\, U^*{}_{rs}{}^i(P^{-1}) = U_{rs}{}^i(P)\, T^*{}_{rp}\, U_{pq}{}^i(P^{-1})\, T_{sq},$$

from (50), and therefore

$$\sum_P \chi^i(P^2) = T^*{}_{rp}\, T_{sq} \sum_P U_{rs}{}^i(P)\, U_{pq}{}^i(P^{-1}) = \frac{N}{n_i} T^*{}_{rp}\, T_{sq}\, \delta_{rq}\, \delta_{sp} \qquad \ldots\ldots(56)$$

$$= \frac{N}{n_i} \text{spur}\,(TT^*) = \pm N, \quad \text{according as } T = \pm \tilde{T}, \qquad \ldots\ldots(57)$$

where we have invoked the generalized orthogonality relations in (56), and relations (52) and $c = \pm 1$, in (57).

The proof that $\mathscr{R}\mathrm{S}^i = \mathrm{S}^i$ implies $\Sigma_P \chi(P^2) = N(-1)^\nu$, follows at once from (57), once we have proved that $\mathscr{R}\mathrm{S}^i = \mathrm{S}^i$ implies

$$T = (-1)^\nu \tilde{T}. \qquad \ldots\ldots(58)$$

To prove (58) we note that the hypothesis $\mathscr{R}\mathrm{S}^i = \mathrm{S}^i$ implies that

$$\mathscr{R}\Psi_r = \Psi_s T_{sr} \quad \text{where} \quad \Psi_s . \mathscr{R}\Psi_r = T_{sr},$$

since $\{\Psi_r\}$ and $\{\mathscr{R}\Psi_r\}$ are orthonormal bases in S^i, related by the unitary matrix T. It follows at once, therefore, from (41), that $\mathscr{R}\mathrm{S}^i = \mathrm{S}^i$ implies

$$T_{rs} = \Psi_r . \mathscr{R}\Psi_s = (-1)^\nu \Psi_s . \mathscr{R}\Psi_r = (-1)^\nu T_{sr};$$

that is $$T = (-1)^\nu \tilde{T}. \qquad \ldots\ldots(59)$$

In certain cases we do not know the explicit character χ^i of S^i at the outset, and it is convenient to calculate Wigner's sum $\Sigma_P \chi^i(P^2)$ in the form

$$\sum_r \sum_P U_{rr}{}^i(P^2).$$

It is to meet this situation that Wigner extends his theorem to include the assertion that $\Sigma_P \chi^i(P^2) = 0$ implies $(\chi^i)^* \neq \chi^i$. In order to prove this assertion, we first prove the converse, namely $(\chi^i)^* \neq \chi^i$ implies $\Sigma_P \chi^i(P^2) = 0$. We can choose a basis in S^i such that the corresponding irreducible representation $\{U^i(P)\}$ is unitary; then

$$\sum_P \chi^i(P^2) = \sum_P U_{rs}{}^i(P)\, U_{sr}{}^i(P) = \sum_P U_{rs}{}^i(P)\, U^*{}_{rs}{}^i(P^{-1}) = 0,$$

from the generalized orthogonality relations, since by hypothesis, $\{U^i(P)\}$ and $\{U^{*i}(P)\}$ are inequivalent irreducible representations.

We can now assert that

(*a*) $\sum_P \chi^i(P^2) = \pm N$ implies χ^i real,

(*b*) $\sum_P \chi^i(P^2) = 0$ implies χ^i complex,

for we have shown that

(*c*) χ^i real implies $\sum_P \chi^i(P^2) = \pm N$,

(*d*) χ^i complex implies $\sum_P \chi^i(P^2) = 0$.

Kramers degeneracy.

If $\mathcal{N}$, the number of electrons in the system, is odd and no external magnetic field is present, we can assert, without further qualification, that the degeneracy of all energy levels is of even order. This assertion is clearly true for those levels for which time-reversal degeneracy exists, so let us consider the case $\mathscr{R}S^i = S^i$. It follows from (59) that $T = -\tilde{T}$, since $\mathcal{N}$ is odd. Now the determinant of a matrix is equal to the determinant of the transposed matrix, therefore

$$\det T = \det \tilde{T} = \det(-T) = (-1)^n \det T.$$

It follows, since $\det T \neq 0$, that the dimensionality n of S^i is even. This assertion is due to Kramers (1930).

Accidental degeneracy.

We now show, if $\mathscr{R}S^i \neq S^i$ and we find $\Sigma_P \chi^i(P^2) = +N(-N)$, for $\mathcal{N}$ even (odd), that the degeneracy cannot be attributed to $\mathscr{R}$. The proof consists of showing that in this case, the space representing the sum of S^i and $\mathscr{R}S^i$ can be decomposed into two complementary sub-spaces, each of which is irreducible under $\{P\}$ and invariant under $\mathscr{R}$. Starting from the given hypothesis,

$$\sum_P \chi^i(P^2) = +N(-N),$$

for $\mathcal{N}$ even (odd), we note that this implies $T = (-1)^{\mathcal{N}} \tilde{T}$ from (57). Furthermore, if $\{\Psi_r\}$ is an orthonormal basis in S^i, then $\{\Phi_s\}$, where

$$\Phi_s = (\mathscr{R}\Psi_r)\, T_{rs}^{-1}, \qquad \ldots\ldots(60)$$

is the isomorphic basis in the space $\mathscr{R}S^i$. Now consider the space S^+ spanned by the orthogonal basis vectors $\{\Psi_s + \Phi_s\}$, and the space S^- spanned by the orthogonal vectors $\{\Psi_s - \Phi_s\}$. These spaces S^+ and S^- are each invariant and irreducible under $\{P\}$, for

$$P(\Psi_s \pm \Phi_s) = (\Psi_r \pm \Phi_r)\, U_{rs}^i(P).$$

They are also each invariant under $\mathscr{R}$, for

$$\begin{aligned}
\mathscr{R}(\Psi_s \pm \Phi_s) &= \mathscr{R}[\Psi_s \pm (\mathscr{R}\Psi_r)\, T_{rs}^{-1}] \quad \text{from (60)},\\
&= \mathscr{R}\Psi_s \pm (-1)^{\mathcal{N}} \Psi_r T^*_{rs}{}^{-1} \quad \text{from (43)},\\
&= \mathscr{R}\Psi_s \pm (-1)^{\mathcal{N}} \Psi_r T_{sr}, \quad \text{since } T \text{ is unitary},\\
&= \mathscr{R}\Psi_s \pm (-1)^{\mathcal{N}} \Psi_r (-1)^{\mathcal{N}} T_{rs}, \quad \text{since } T = (-1)^{\mathcal{N}} \tilde{T}\\
&= (\Phi_r \pm \Psi_r)\, T_{rs}, \quad \text{from (60)}.
\end{aligned}$$

The sum of these spaces S^+ and S^- is equal to the eigenvector sub-space $S(E)$ formed from the sum of the spaces S^i and $\mathscr{R}S^i$ given in the initial hypothesis. We now show that S^+ and S^- are complementary sub-spaces in $S(E)$; it then follows, since S^+ and S^- are each invariant under $\mathscr{R}$ and $\{P\}$, that the degeneracy implied by the hypothesis $\mathscr{R}S^i \neq S^i$ cannot be attributed to $\mathscr{R}$.

In order to prove that S^+ and S^- are complementary sub-spaces in $S(E)$, we observe, since S^+ and S^- are irreducible under $\{P\}$, that S^+ and S^- are either identical or contain no non-zero vector in common. But they are not identical, for this would imply, for example, that

$$(\Psi_s - \Phi_s) = \sum_r (\Psi_r + \Phi_r)\, c_{rs},$$

$$= (\Psi_s + \Phi_s)c,$$

since $$(\Psi_s - \Phi_s).(\Psi_r + \Phi_r) = 0 \quad \text{if } r \neq s,$$

and $(\Psi_s - \Phi_s).(\Psi_s + \Phi_s)$ is independent of s, from the properties of such bases (Part 1, §6). But these relations imply $\mathscr{R}S^i = S^i$, contrary to the hypothesis $\mathscr{R}S^i \neq S^i$, for they imply

$$\Phi_s = \left(\frac{1-c}{1+c}\right)\Psi_s,$$

and therefore, from (60), $$\mathscr{R}\Psi_r = \left(\frac{1-c}{1+c}\right)\Psi_s T_{sr}.$$

If this hypothetical degeneracy did occur in practice, we should conclude that the symmetry group $\{P\}$ was incomplete. If the degeneracy appeared in a calculation based on a model, it could be accidental, due to an approximation or a special choice of values for disposable parameters.

§4. Group Theory and the Many-electron Problem

We are familiar with the fact that the wave functions of a many-electron system are antisymmetric in the electron labels. It has been common practice for many years in solid-state physics to take account of this fact by working in terms of determinantal wave functions. It is possible, however, to take account of the Pauli principle by a group-theoretical method, which, in the absence of spin–orbit coupling terms in the Hamiltonian of the model, is potentially much more powerful than the determinantal approach.

This group-theoretical method has been known almost as long as the Pauli principle, but has had to await the coming of the automatic computer with a large fast store, before coming into its own. The idea underlying the method is, given a product space $S = S^i \times S^j$ formed from two spaces S^i and S^j, each of which affords an irreducible representation of the general permutation group Π_n, that the antisymmetric representation of Π_n occurs in the decomposition of the product space S, once and once only, if, and only if, S^i and S^j afford adjoint (associated) representations of Π_n (Part 1, §6). If, on the other hand, S^i and S^j are not adjoint, then the antisymmetric representation of Π_n does not occur at all. The proof of these assertions follows very simply from the theorems given in Part 1, §6. Let $\chi^i \equiv \{\chi^i(P)\}$, $\chi^j \equiv \{\chi^j(P)\}$ denote the characters of S^i and S^j respectively, and $\{\zeta(P)\}$ the character of the antisymmetric representation of Π_n. The character of

$S = S^i \times S^j$ is then $\{\chi^i(P)\chi^j(P)\}$. The number of times $\{\zeta(P)\}$ appears in the decomposition of the character of S is

$$c = \frac{1}{n!}\sum_P \zeta(P^{-1})\chi^i(P)\chi^j(P). \qquad \ldots\ldots(61)$$

If χ^i and χ^j are adjoint characters, then by definition $\chi^j(P) = \zeta(P)\chi^i(P)$, and

$$c = \frac{1}{n!}\sum_P \zeta(P^{-1})\,\zeta(P)[\chi^i(P)]^2. \qquad \ldots\ldots(62)$$

Now it is a property of the permutation groups that a permutation P and its reciprocal P^{-1} are members of the same class (P and P^{-1} have the same cycle structure while the cycle structure determines the class (Ledermann 1949)). It follows that $\chi^i(P) = \chi^i(P^{-1})$, and so $[\chi^i(P)]^2 = \chi^i(P)\chi^i(P^{-1})$. Furthermore, $\zeta(P^{-1})\,\zeta(P) = 1$. Making these substitutions in (62), we obtain

$$c = \frac{1}{n!}\sum_P \chi^i(P)\chi^i(P^{-1}) = 1, \qquad \ldots\ldots(63)$$

since χ^i is an irreducible character of Π_n. If χ^i and χ^j are not adjoint, then the sum in (61) is zero, from the character orthogonality relations, since $\{\chi^i(P)\}$ and $\{\zeta(P)\chi^j(P)\}$ are then inequivalent irreducible characters. Summarizing

$$\left.\begin{matrix} c = 1 \\ c = 0 \end{matrix}\right\} \text{ if } \chi^i \text{ and } \chi^j \left\{\begin{matrix} \text{are} \\ \text{are not} \end{matrix}\right\} \text{ adjoint characters of } \Pi_n. \qquad \ldots\ldots(64)$$

We exploit this theorem as follows.

The state space of our n-electron model is a product of the n one-electron spaces, each of which is a product of a function space with a two-dimensional spin space. We can, however, regard this state space as the product of two spaces, S_n and O_n, where S_n is the product of the n two-dimensional spin spaces, and O_n is the product of the n function spaces, one for each electron. We require all those vectors in the space $S_n \times O_n$ which afford the antisymmetric representation $\{\zeta(P)\}$ of Π_n. These are the vectors which satisfy the Pauli principle, and we shall call them Pauli vectors.

Recalling our definition of permutation operations on product spaces in Part 1, §3, we note that each of the spaces S_n and O_n is invariant under the group of permutations Π_n on the n labels of the electrons. The space S_n will be of 2^n dimensions. We can decompose the space O_n into finite sub-spaces, each of which is generated from a single product of n functions (one function to each electron but not necessarily all different), by the operations of Π_n. The maximum dimensionality of such a sub-space is then $n!$, attained when all the one-electron functions in the initial vector are different. These sub-spaces are invariant under Π_n by their method of construction. As an economy in notation we now let O_n denote one of these sub-spaces. We denote the vector from which O_n is generated by the symbol

$$F(\ldots),$$

which represents the ordered product of n one-electron orbital functions. We call $F(\ldots)$ a Hartree orbital product, since these products form the basis of Hartree's self-consistent field calculations.

Now we can decompose the spin space S_n into sub-spaces $S(p, \lambda)$, irreducible under Π_n, where λ labels the irreducible character of the sub-space, and the label p distinguishes between the multiple occurrences of sub-spaces with the same character λ. Similarly, we decompose O_n into sub-spaces $O(k, \mu)$, where μ labels the character, and k distinguishes between the multiple occurrences of a given character. The space $S_n \times O_n$ can now be expressed as a sum of sub-spaces of the from $S(p, \lambda) \times O(k, \mu)$. We require all the Pauli vectors contained in each of these sub-spaces. We know however, from (64), that there are no Pauli vectors contained in such a sub-space if the character χ^μ is not adjoint to the character χ^λ, while if the characters are adjoint, the sub-space $S(p, \lambda) \times O(k, \mu)$ contains one Pauli vector and one only. It follows that we have to consider only those sub-spaces $O(k, \mu)$ of O_n which are adjoint to sub-spaces $S(p, \lambda)$ occurring in the decomposition of S_n. We label the Pauli vectors constructed in this way with the corresponding label λ (further distinguishing labels will appear later).

The significance of this method of choosing Pauli vectors is that the matrix elements of the Hamiltonian of the system, taken between Pauli vectors with different λ labels, are zero, if the Hamiltonian does not contain spin–orbit coupling terms. For example, the matrix elements of the electrostatic electron–electron interaction are zero. It follows that, to this approximation, the eigenvalues of the Hamiltonian can be distinguished by the labels λ, which appear in the decomposition of the spin space S_n, although at the outset it appeared that only the antisymmetric character of Π_n was relevant to the problem. It is possible, in principle, to construct Pauli vectors with similar properties, by starting from determinantal wave functions. This however, involves forming linear combinations of determinants, which are eigenvectors of the square of the total spin of the system, a calculation only tractable for a relatively small number of electrons.

From a computational point of view, these matrix elements of the Hamiltonian, with respect to Pauli vectors, form the raw material for the calculation of eigenvalues by the method described in Part 1, § 7.

The calculation of Pauli vectors.

We turn now to the problem of the explicit construction of our Pauli vectors. We shall quote a number of theorems without proof. These theorems can be found, more or less disguised, in the literature on the theory of the permutation groups (Littlewood 1950, Murnaghan 1938, Rutherford 1948, Weyl 1939). The first theorem we use is that, to every irreducible character of the permutation group Π_n, there corresponds a real unitary matrix representation of Π_n.

Let $\{\phi_r\}$ denote an orthonormal basis in the sub-space $S(p, \lambda)$, such that $P\phi_s = \Sigma_r \phi_r U_{rs}^\lambda(P)$, where $\{U^\lambda(P)\}$ is a real unitary matrix representation of Π_n, with character χ^λ. Similarly let $\{\psi_r\}$ denote an orthonormal basis in an adjoint sub-space $O(k, \bar{\lambda})$, where $\bar{\lambda}$ denotes the adjoint of λ, with the property

$$P\psi_s = \sum_r \psi_r \zeta(P) U_{rs}^{\bar{\lambda}}(P).$$

Consider the vector $\phi_r\psi_r$ in $S(p, \lambda) \times O(k, \bar{\lambda})$. In order to construct a Pauli vector from the vector $\phi_r\psi_r$, we use the projection operator technique described in Part 1, § 7.

The operator corresponding to the antisymmetric representation is

$$e = \frac{1}{n!}\sum_P \zeta(P)P. \qquad \ldots\ldots(65)$$

Then
$$\left.\begin{aligned} e(\phi_r\psi_r) &= \frac{1}{n!}\sum_P \zeta(P)(P\phi_r)(P\psi_r) \\ &= \frac{1}{n!}\sum_{s,l}\phi_s\psi_l\sum_P U_{sr}{}^{\lambda}(P)\,U_{lr}{}^{\lambda}(P) \\ &= \frac{1}{n_\lambda}\sum_s \phi_s\psi_s, \end{aligned}\right\} \qquad \ldots\ldots(66)$$

from the generalized orthogonality relations, since the representation $\{U^\lambda(P)\}$ is real. The normalized Pauli vector is $\sqrt{n_\lambda}\, e\, k(\phi_r\psi_r)$, since $\{\phi_r\}$ and $\{\psi_r\}$ are orthonormal bases. Since we know that there is only one Pauli vector to be found in the product space $S(p,\lambda)\times O(k,\tilde{\lambda})$, we have completed our search at the first attempt!

Matrix elements.

We now show that the matrix elements of the Hamiltonian H, taken between Pauli vectors

$$\Psi(\lambda) = \sqrt{n_\lambda}\, e(\phi_r{}^{\lambda}\psi_r{}^{\tilde{\lambda}}) = \frac{1}{\sqrt{n_\lambda}}\sum_s \phi_s{}^{\lambda}\psi_s{}^{\tilde{\lambda}},$$

and
$$\Psi(\mu) = \sqrt{n_\mu}\, e(\phi_{p'}{}^{\mu}\psi_{p'}{}^{\tilde{\mu}}) = \frac{1}{\sqrt{n_\mu}}\sum_q \phi_q{}^{\mu}\psi_q{}^{\tilde{\mu}},$$

is zero for $\lambda \neq \mu$, if H is of the form $H = H' + H''$, where H' is symmetric in the electron space coordinates, and H'' is symmetric in the electron spin coordinates, there being no spin–orbit cross-terms (that is, $\mathscr{S}$ is a good quantum number). Consider first

$$\left.\begin{aligned} \Psi(\lambda).H'\Psi(\mu) &= (n_\lambda n_\mu)^{1/2}\, e(\phi_r{}^{\lambda}\psi_r{}^{\tilde{\lambda}}).H'\, e(\phi_{p'}{}^{\mu}\psi_{p'}{}^{\tilde{\mu}}) \\ &= (n_\lambda n_\mu)^{1/2}\,\phi_r{}^{\lambda}\psi_r{}^{\tilde{\lambda}}.H'\, e(\phi_{p'}{}^{\mu}\psi_{p'}{}^{\tilde{\mu}}), \end{aligned}\right\} \qquad \ldots\ldots(67)$$

where we have made use of the fact that the operator e is Hermitian, commutes with H', and is idempotent, that is $e^2 = e$. Substituting for $e(\phi_{p'}{}^{\mu}\psi_{p'}{}^{\tilde{\mu}})$ from (66) in (67)

$$\Psi(\lambda).H'\Psi(\mu) = \left(\frac{n_\lambda}{n_\mu}\right)^{1/2}\phi_r{}^{\lambda}\psi_r{}^{\tilde{\lambda}}H'\sum_s \phi_s{}^{\mu}\psi_s{}^{\tilde{\mu}} \qquad \ldots\ldots(68)$$

$$= \left(\frac{n_\lambda}{n_\mu}\right)^{1/2}\sum_s [\psi_r{}^{\tilde{\lambda}}.H'\psi_s{}^{\tilde{\mu}}]\,[\phi_r{}^{\lambda}.\phi_s{}^{\mu}], \qquad \ldots\ldots(69)$$

since H' is a function of the electron space coordinates alone. But $[\phi_r{}^{\lambda}.\phi_s{}^{\mu}] = 0$ if $\lambda \neq \mu$, therefore

$$\Psi(\lambda).H'\Psi(\mu) = 0, \quad \text{if } \lambda \neq \mu. \qquad \ldots\ldots(70)$$

Similarly
$$\Psi(\lambda).H''\Psi(\mu) = 0, \quad \text{if } \lambda \neq \mu. \qquad \ldots\ldots(71)$$

In order to evaluate the matrix elements of H in the case $\mu = \lambda$, we require an elaboration of our notation, to take account of the fact that, in general, more than

one Pauli vector can be constructed corresponding to a given choice of λ. We now denote the basis in the sub-space $S(p,\lambda)$ of spin space by $\{\phi(p,r,\lambda)\}$. We have then from (69), for the case $\mu = \lambda$,

$$\Psi(p,j,\lambda).H'\Psi(q,k,\lambda) = [\psi(j,r,\bar{\lambda}).H'\psi(k,r,\bar{\lambda})]\,[\phi(p,r,\lambda).\phi(q,r,\lambda)], \quad \ldots\ldots(72)$$

where we have taken account of the fact that

$$\phi(p,r,\lambda).\phi(q,s,\lambda) = 0 \quad \text{if} \quad r \neq s.$$

Similarly $$\Psi(p,j,\lambda).H''\Psi(q,k,\lambda) = [\psi(j,r,\bar{\lambda}).\psi(k,r,\bar{\lambda})]\,[\phi(p,r,\lambda).H''\phi(q,r,\lambda)]. \quad \ldots\ldots(73)$$

We now consider in more detail the choice of the sub-spaces $S(p,\lambda)$ and their corresponding properties.

Properties of the spin sub-spaces $S(p,\lambda)$.

We take as our basis vectors in the spin space S_n, the 2^n Hartree spin products. Each of these vectors is a product of n_α one-electron spin vectors α and n_β spin vectors β, using a familiar notation, where $n_\alpha + n_\beta = n$, and α and β are orthogonal and normalized. Starting from the basic vector

$$\underbrace{\alpha\alpha\ldots\alpha}_{n_\alpha}\ \underbrace{\beta\beta\ldots\beta}_{n_\beta}$$

corresponding to a given choice of the partition $n_\alpha + n_\beta = n$, we can generate a sub-space of S_n, by operating on this particular vector with all the permutation operations contained in Π_n. This space $S(n_\alpha, n_\beta)$ will be of dimensions $n!/n_\alpha!\,n_\beta!$. We now assert that in the decomposition of the character of the sub-space $S(n_\alpha, n_\beta)$, if an irreducible character of Π_n occurs at all, then it occurs exactly once. This remarkable theorem, which we do not attempt to prove, is a particular case of a more general theorem, relating the permutation groups with groups known as the full linear groups (Murnaghan 1938). It follows at once from the foregoing assertion, if we select our sub-spaces $S(p,\lambda)$ as sub-spaces of the spaces $S(n_\alpha, n_\beta)$, that we can identify the label p with a partition of n. We now observe, using this definition of p, that two sub-spaces $S(p,\lambda)$, $S(q,\lambda)$, are mutually orthogonal if p and q correspond to different partitions of n. This result follows at once from the chosen orthogonality of the one-electron spin vectors α and β. Returning to equation (72), this mutual orthogonality implies

$$[\phi(p,r,\lambda).\phi(q,r,\lambda)] = \delta_{pq} \quad \ldots\ldots(74)$$

with the result

$$\Psi(p,j,\lambda).\mathrm{H}'\Psi(q,k,\lambda) = [\psi(j,r,\bar{\lambda}).H'\psi(k,r,\bar{\lambda})]\,\delta_{pq}. \quad \ldots\ldots(75)$$

Configuration interaction.

In general, the vectors $\psi(j,r,\bar{\lambda})$ and $\psi(k.r,\bar{\lambda})$ in (75) may come from different orbital spaces O_n, generated from orbital Hartree products constructed from different sets of orbital functions. The corresponding matrix elements of the

Hamiltonian are said to represent configuration interaction, a configuration being defined by an orbital Hartree product.

The decomposition of the character of S_n.

It is evident that in order to use this method, we need to know at the outset,

(i) the decomposition of the character of a chosen sub-space S_n,

(ii) the decomposition of the character of a chosen sub-space O_n.

The theory of permutation groups shows that the irreducible characters of Π_n can be placed in one-to-one correspondence with the partitions of n. It can be shown moreover, that the irreducible characters of Π_n that appear in the decomposition of the character of S_n correspond to two element partitions. We denote these two element partitions and the corresponding characters by the same symbol $[n-r, r]$, where r ranges from 0 to $(n-1)/2$ if n is odd, and from 0 to $n/2$ if n is even. Furthermore, the character $[n-r, r]$ occurs $n-2r+1$ times in the decomposition of S_n; that is

$$\text{character of } S_n = \sum_r (n-2r+1)[n-r, r]. \qquad \ldots\ldots(76)$$

This decomposition is all we need to know about S_n in the evaluation of the matrix elements of H′ in (75). An assertion equivalent to (76) is that the character of the sub-space $S(n_\alpha, n_\beta)$ of S_n, corresponding to the partition $n_\alpha + n_\beta = n$, is the sum of the irreducible characters

$$[n], [n-1, 1], \ldots [n_\alpha + 1, n_\beta - 1], [n_\alpha, n_\beta],$$

each occurring once. This is a more explicit statement of the theorem quoted in our discussion of the label p in our notation $S(p, \lambda)$. We can write symbolically

$$S(n_\alpha, n_\beta) = \sum_\lambda S(p, \lambda), \quad p \equiv (n_\alpha, n_\beta) \qquad \ldots\ldots(77)$$

where λ runs from $[n]$ to $[n_\alpha, n_\beta]$. The reader may find it helpful to analyse the eight-dimensional spin space of a three-electron system with respect to the permutation group Π_3. This space was discussed as an example in Part 1.

Character tables of Π_n.

We note here that the character adjoint to $[n-r, r]$ corresponds to the partition

$$[2^r, 1^{n-2r}]. \qquad \ldots\ldots(78)$$

The character of the unit representation of Π_n corresponds to the partition $[n]$. The character of the antisymmetric representation of Π_n corresponds to the partition $[1^n]$. The dimensionality of an irreducible representation of Π_n with character $[n-r, r]$ is

$$\frac{n!\,(n-2r+1)}{(n-r+1)!\,r!} \qquad \ldots\ldots(79)$$

for $r \geqslant 1$. For $r = 0$ the dimensionality is unity. Tables of characters for $\Pi_n (n = 2 \ldots 10)$, can be found in Littlewood (1950). In using these character

tables, the reader should note that the classes of Π_n are characterized by partitions of n, for the class to which an element of Π_n belongs is determined by the cycle structure of the element, while the different possible cycle structures are characterized by partitions of n. (See, for example, Ledermann (1949), Margenau and Murphy (1943), Rutherford (1948)). To avoid confusion of these two uses of partitions in the character table of Π_n, it is conventional to use distinct symbols for the partitions labelling characters and the partitions labelling classes.

The total spin quantum number $\mathscr{S}$.

It can be deduced that for a given value of n, there is a one-to-one correspondence between the character label $[n-r, r]$ of Π_n, and the total spin quantum number $\mathscr{S}$, where $\mathscr{S}(\mathscr{S}+1)$ is an eigenvalue of the operator

$$\left[\sum_{i=1}^{i=n} \tfrac{1}{2}\sigma_x(i)\right]^2 + \left[\sum_{i=1}^{i=n} \tfrac{1}{2}\sigma_y(i)\right]^2 + \left[\sum_{i=1}^{i=n} \tfrac{1}{2}\sigma_z(i)\right]^2$$

representing the square of the total spin of the system. The $(1+2\mathscr{S})$-fold degeneracy of the eigenvalue $\mathscr{S}(\mathscr{S}+1)$ corresponds in fact to the $(n-2r+1)$-fold occurrence of the character $[n-r, r]$ in (76), so that

$$\mathscr{S} = \tfrac{1}{2}n - r.$$

It follows that $\mathscr{S}$ ranges from $\frac{1}{2}n$ to zero if the number of electrons n is even, and from $\frac{1}{2}n$ to $\frac{1}{2}$ if n is odd, a well-known result.

The Heitler–London model.

There is one form of orbital space O_n, which we can decompose at once with respect to Π_n. This is the case where the initial orbital Hartree product vector $F(\ldots)$ is a product of n different one-electron orbital functions. It follows at once that the corresponding space O_n affords the regular representation of Π_n. A model based on this choice of Hartree product is a Heitler–London model, for the electron orbitals are only singly occupied. Although the physical relevance of a Heitler–London model is limited, such a model does provide an excellent first exercise in the group-theoretical method.

In Part 1, § 7, we showed how a space affording the regular representation of a finite group can be decomposed into irreducible sub-spaces by use of the projection operators contained in the group algebra. In our case we must choose projection operators

$$\epsilon_{rs}^{\lambda} = \frac{n_\lambda}{n!} \sum_P U_{sr}^{\lambda}(P^{-1})\, P, \quad P \text{ operating on } O_n,$$

corresponding to real unitary representations of Π_n, if (66) is to be valid. Let us consider the evaluation of the matrix element $[\psi(j, r, \lambda).H'\psi(k, r, \lambda)]$ for a Heitler–London model, limited to a single configuration, that is, based on a single orbital Hartree product. In order to simplify the algebra, we assume that the n different functions forming the orbital Hartree product $F(\ldots)$, are themselves members of an orthonormal set. With this assumption $(F.PF)$, $P \in \Pi_n$, is zero if P is not equal to the unit permutation I. We can then write for our normalized

vectors $\psi(j,r,\tilde{\lambda})$ and $\psi(k,r,\tilde{\lambda})$,

$$\left.\begin{aligned}\psi(j,r,\tilde{\lambda}) &= \left(\frac{n!}{n_\mu}\right)^{\frac{1}{2}} \epsilon_{rs}{}^{\mu} F,\\ \psi(k,r,\tilde{\lambda}) &= \left(\frac{n!}{n_\mu}\right)^{\frac{1}{2}} \epsilon_{rt}{}^{\mu} F,\end{aligned}\right\} \qquad \ldots\ldots(80)$$

where the labels s and t replace j and k, and μ denotes one of the adjoint partitions (78). These vectors are normalized since, for example,

$$(\epsilon_{rs}{}^{\mu} F).(\epsilon_{rs}{}^{\mu} F) = (F.\epsilon_{ss}{}^{\mu} F) = \frac{n_\mu}{n!}\sum_P (F.PF)\, U_{ss}{}^{\mu}(P^{-1}) = \frac{n_\mu}{n!},$$

using the condition $(F.PF) = 0$ if $P \neq I$, and the multiplication rules for the projection operators. With this substitution, we have

$$\begin{aligned}[\psi(s,r,\mu).H'\psi(t,r,\mu)] &= \frac{n!}{n_\mu}(\epsilon_{rs}{}^{\mu} F).H'(\epsilon_{rt}{}^{\mu} F)\\ &= \sum_P (F.H'PF)\, U_{st}{}^{\mu}(P). \qquad \ldots\ldots(81)\end{aligned}$$

Now suppose that the Hamiltonian H' is of the form $(H+\Sigma_{i<j} V_{ij})$, where V_{ij} denotes the interaction between the ith and jth electrons, and $H = \Sigma_i H_i$ is a sum over one-electron operators H_i. Then $(F.H_iPF) = 0$ if $P \neq I$, while

$$(F.H_iF) = (\psi_i H_i \psi_i),$$

where ψ_i is the orbital function allocated to the ith electron in the Hartree product F. Now consider the terms $(F.V_{ij}PF)$. It follows from the structure of F that $(F.V_{ij}PF)$ is zero, unless $P = I$ or $P = P_{ij}$, the permutation that interchanges the labels on the ith and jth electrons. For this form of Hamiltonian then

$$\begin{aligned}[\psi(s,r,\mu).H'\psi(t,r,\mu)] &= \sum_i (\psi_i.H_i\psi_i)\,\delta_{st} + \sum_{i<j}(F.V_{ij}F)\,\delta_{st}\\ &\quad + \sum_{i<j}(F.V_{ij}P_{ij}F)\,U_{st}{}^{\mu}(P_{ij}), \qquad \ldots\ldots(82)\end{aligned}$$

since $U_{st}{}^{\mu}(I) = \delta_{st}$ in (81). We note that the only non-zero off-diagonal matrix elements of H' are the 'exchange' terms generated by the permutations P_{ij}. Furthermore, the contribution from the first two terms on the right-hand side of (82) is the same for all the diagonal elements. In a calculation of eigenvalues of H' from the secular equation

$$\det[\psi(s,r,\mu).H'\psi(t,r,\mu) - W^{\mu}\psi(s,r,\mu).\psi(t,r,\mu)] = 0, \quad \ldots\ldots(83)$$

these terms only affect the zero of the energy scale. The relative eigenvalues are given by the equation

$$\det(H_{st} - W^{\mu}\delta_{st}) = 0,$$

where

$$\begin{aligned}H_{st} &= \sum_{i<j}(F.V_{ij}P_{ij}F)\,U_{st}{}^{\mu}(P_{ij})\\ &= -\sum_{i<j}(F.V_{ij}P_{ij}F)\,U_{st}{}^{\lambda}(P_{ij}),\ \tilde{\lambda} = \mu. \qquad \ldots\ldots(84)\end{aligned}$$

This result shows explicitly how the elements of the secular determinant are related to the matrix elements of the representation λ of Π_n. The order of this determinant is given by (79).

The Dirac equivalent Hamiltonian.

Let P_{ij} denote the permutation operation on our spin space S_n, which interchanges the labels of the ith and jth electrons, and consider the eigenvalues of the operator

$$D = -\sum_{i<j}(F.V_{ij}P_{ij}F)P_{ij}. \qquad \ldots\ldots(85)$$

From (77) we can decompose S_n in the form

$$S_n = \Sigma S(n_\alpha, n_\beta) = \sum_{p,\lambda} S(p, \lambda). \qquad \ldots\ldots(86)$$

We can choose a basis in $S(p, \lambda)$ such that the corresponding matrix representative of the permutation P_{ij} is the matrix $U^\lambda(P_{ij})$, chosen in the construction of (84). It follows at once that the eigenvalues of the operator D in $S(p, \lambda)$ are identical to the eigenvalues of the matrix $\|H_{\alpha\beta}\|$ of (84). Extending these ideas to the whole space $S_n = \Sigma_{p,\lambda} S(p, \lambda)$, we can assert that the eigenvalues of the operator D in the whole space S_n are the eigenvalues of the Heitler–London model. The operator D is the Dirac equivalent Hamiltonian (Dirac 1947).

If we wish to disguise the group-theoretical background of the operator D, we can replace the operator P_{ij} by the exactly equivalent operator on S_n,

$$P_{ij} \equiv \tfrac{1}{2}[1 + \boldsymbol{\sigma}(i).\boldsymbol{\sigma}(j)];$$

then

$$D = -\tfrac{1}{2}\sum_{i<j} J_{ij}[1 + \boldsymbol{\sigma}(i).\boldsymbol{\sigma}(j)], \qquad \ldots\ldots(87)$$

where

$$J_{ij} = (F.V_{ij}P_{ij}F).$$

The advantage of working from (84) instead of (87) is, of course, that in (84) our secular determinant only relates to eigenvalues corresponding to a single irreducible character χ^λ; whereas in the solution of (87) for a sub-space $S(n_\alpha, n_\beta)$, the corresponding secular equation includes eigenvalues from all the sub-spaces $S(p, \lambda)$ occurring in the decomposition formula (77). This fact has been exploited by, for example, Mattheiss (1959).

Doubly occupied orbitals.

In practice the assumption of single occupation, basic to the Heitler–London model, is generally too restrictive, and we have to consider orbital spaces O_n generated from Hartree products $F(\ldots)$, in which the orbital functions are not all different. No given orbital may occur more than twice in $F(\ldots)$, of course, as we know from a determinantal approach. In fact, if an orbital does occur more than twice, then it is found that the character of the corresponding space O_n does not contain, on decomposition, any irreducible character of the required form (78).

An extensive literature exists on the subject of doubly occupied orbitals. Various schemes have been proposed, each with its own peculiar merits, but none of them can be regarded as the last word on the subject. We shall content ourselves here with certain comments on the problem and refer the reader to the articles by

Yamanouchi (1936), Corson (1951), McWeeny (1955), Koster (1956). There is no difficulty, in principle, in considering doubly occupied orbitals, for we can still use the projection operator technique to decompose an orbital space O_n, constructed from a Hartree product containing doubly occupied orbitals, provided we recognize linear dependence, as described in Part 1, §7. In order to decide which projection operators to construct, we need to know which irreducible characters $[2^r, 1^{n-2r}]$ appear in the decomposition of the character of a given space O_n. We can obtain this information in fact, without first calculating the character of O_n, by making use of a rule based on a theorem on cosets proved in the Appendix. Before we invoke this rule, we observe that the Hartree product F, from which O_n is generated, is itself invariant under a sub-group of Π_n, namely, that sub-group generated by the permutations that interchange the labels of the electrons occupying the doubly occupied orbitals. We denote this sub-group by G_F. For example, if none of the orbitals in F are doubly occupied, the corresponding group G_F contains just the unit permutation. If the number of pairs of doubly occupied orbitals is m, then the number of elementals in G_F is 2^m.

Now any given irreducible representation λ of Π_n contains a representation, say Γ, of G_F. In general this representation Γ of G_F is reducible. In particular, given the character χ^λ of Π_n, we can calculate how many times, say C^λ, the unit representation of G_F occurs in the decomposition of Γ. To do this, we do not need to know the complete character χ^λ, it is sufficient to know the spurs of the permutations common to G_F and Π_n. The rule, which we now invoke, states that C^λ defined above is equal to the number of times the character χ^λ occurs in the decomposition of the character of the space O_N, corresponding to the Hartree product defining G_F. The existence of this rule is due to the way the space O_n is generated from the Hartree product F by the operations of Π_n. As an example of this rule, we can take F to contain no doubly occupied orbitals, then the corresponding group G_F is just the unit permutation. The number of times the unit representation of G_F occurs in any irreducible representation λ of Π_n is just n_λ, that is, using the rule above, the character χ^λ of Π_n occurs n_λ times in the decomposition of the character of the chosen O_n. This of course is what we expect, since by hypothesis, the given space O_n affords the regular representation of Π_n.

It is possible to simplify the projection operators required in the case of doubly occupied orbitals, by appealing to the theorems on cosets proved in the Appendix. Assuming that the reader has studied these theorems, we observe that each of the $n!$ permutations of Π_n, which appear in a projection operator $\epsilon_{rs}{}^\lambda$, can be expressed as the product of an element S of the sub-group G_F with a left coset-representative Q of Π_n with respect to G_F; that is, $P = QS$. Since the given Hartree product F is invariant under the elements S of G_F, we have $PF = QF$, and only the permutation operators Q occur explicitly in the expression $\epsilon_{rs}{}^\lambda F$. If the order of the sub-group G_F is 2^m, the number of left coset-representatives Q is $n!/2^m$.

Space symmetry.

The symmetry group of the Hamiltonian of an n-electron system is the direct product of the group of permutations Π_n on the particle labels, with the space-symmetry group $G \equiv \{g\}$ of the system. If we operate on a given orbital

Hartree product F with one of the space-symmetry operations g, the result gF can be expressed as a linear combination of orbital Hartree products. Starting from a given orbital Hartree product F, we can generate the space spanned by the set of vectors $\{gPF\}$, where $g \in G$ and $P \in \Pi_n$. By construction, this space is invariant under the complete symmetry group $G \times \Pi_n$. We can now decompose this space with respect to $G \times \Pi_n$, in the form described at the end of Part 1, §6. The basis vectors are then of the form

$$\psi(\eta : r, \mu : \alpha, j), \qquad \ldots\ldots(88)$$

where the symbols α and j label ψ relative to G, and η is necessary to distinguish between isomorphic bases. Provided these vectors $\psi(\eta : r, \mu : \alpha, j)$ are normalized, we can use them in (66), for the construction of Pauli vectors. The labelling of these Pauli vectors takes the form

$$\Psi(\eta : p, \mu, \alpha, j). \qquad \ldots\ldots(89)$$

Equation (75) for the matrix elements of the space part of the Hamiltonian now takes the form

$$\Psi(\xi : p, \lambda, \alpha, j).H'\Psi(\eta : q, \lambda, \beta, k) = [\psi(\xi : r, \tilde{\lambda}, \alpha, j).H'\psi(\eta : r, \tilde{\lambda}, \beta, k)\,\delta_{pq}\,\delta_{\alpha\beta}\,\delta_{jk} \qquad \ldots\ldots(90)$$

where we have taken account of the orthogonality properties of the basis vectors $\psi(\eta : r, \mu, \alpha, j)$.

Irreducible representations of Π_n.

The procedure for calculating matrix elements outlined in this section requires a method for generating the real unitary irreducible representations of Π_n, suitable for coding as a programme for a computer. The usual methods for generating these representations are unsuitable for stored programme computation, since they involve non-arithmetic operations. To overcome this difficulty, new methods are being devised (Gabriel 1960, Mattheiss 1959).

§5. Irreducible Representations of Space Groups

In §1 of Part 2 we outlined the definition of the geometrical crystal space groups, in terms of groups of operators of the general form $(\alpha|\mathbf{v}(\alpha)+\mathbf{n})$. We then defined corresponding substitutional operations on functions, denoted by the symbol $[\alpha|\mathbf{v}(\alpha)+\mathbf{n}]$. In this language we say that a potential energy function $V(\mathbf{X})$ has the symmetry of the geometrical space group $\{(\alpha|\mathbf{v}(\alpha)+\mathbf{n})\}$, if for all operations $[\alpha|\mathbf{v}(\alpha)+\mathbf{n}]$ in the substitutional space group, the relation

$$[\alpha|\mathbf{v}(\alpha)+\mathbf{n}]\,V \equiv V \qquad \ldots\ldots(91)$$

is satisfied. Now in §2 of Part 2 we discussed the symmetry operations induced on the electron state-vectors by substitutional operations on the Dirac equation of motion. It follows, if the potential energy term in the Dirac Hamiltonian is invariant under the substitutional space group $\{[\alpha|\mathbf{v}(\alpha)+\mathbf{n}]\}$, that we have to consider a third kind of space group, whose elements are the symmetry operators defined in equation (17). In order to economize in notation, we denote these symmetry operators by the same symbols $[\alpha|\mathbf{v}(\alpha)+\mathbf{n}]$, on the understanding that

in the multiplication rule

$$[\alpha|\mathbf{v}(\alpha)+\mathbf{n}]\,[\beta|\mathbf{v}(\beta)+\mathbf{m}] = [\alpha\beta|\mathbf{v}(\alpha)+n+\alpha\mathbf{v}(\beta)+\alpha\mathbf{m}], \quad \ldots\ldots(92)$$

the product $\alpha\beta$ is an element of the double point group $\{P\}$, defined by equations (25), if we work in terms of the two-component Pauli formalism. On the other hand, the vectors $\alpha\,\mathbf{v}(\beta)$ and $\alpha\,\mathbf{m}$ in (92) are to be evaluated by interpreting α as a geometrical operator in the position space of the system. With these conventions, the set $\{[\alpha|\mathbf{v}(\alpha)+\mathbf{n}]\}$ is the double space group induced by the geometrical space group $\{(\alpha|\mathbf{v}(\alpha)+\mathbf{n})\}$.

Many-electron systems.

Our formalism can be extended at once to an $\mathcal{N}$-electron system, provided we take account of the distinction between the two cases $\mathcal{N}$ even, $\mathcal{N}$ odd. Recalling the discussion at the end of Part 2, § 2, if $\mathcal{N}$ is even, the only irreducible representations of the double space group that can appear are those representations which are identical with the representations of the substitutional space group. If $\mathcal{N}$ is odd, the only representations of the double space-group that can appear are those irreducible representations which are *not* simultaneously irreducible representations of the corresponding substitutional space groups.

If we use a model in which spin–orbit coupling is neglected, the spatial symmetry group of the model is the product of the substitutional space group with the spin–rotation group, independent of whether $\mathcal{N}$ is even or odd. The irreducible representations of the double space group are irrelevant in this case.

Cyclic boundary conditions.

In order to represent an infinite periodic crystal, built up, we assume, from electrons and atomic nuclei, we use a finite quantum-mechanical model, constrained by cyclic boundary conditions defined by a super-lattice. We can interpret these boundary conditions physically if we choose, or we can regard them as mathematical constraints, which are effectively removed in the limit of large size of the unit cell of the super-lattice. We shall adopt the latter point of view, and impose a modulo condition on the crystal lattice vectors $\mathbf{n}$ in the symmetry operators $[\alpha|\mathbf{v}(\alpha)+\mathbf{n}]$, of the form

$$\mathbf{n} \text{ and } \mathbf{m} \text{ are equivalent if } \mathbf{n}-\mathbf{m} = \mathbf{N} \qquad \ldots\ldots(93)$$

where $\mathbf{N}$ is a vector in the super-lattice defining the cyclic boundary conditions.

With this modulo condition, the space group $\{[\alpha\,\mathbf{v}|(\alpha)+\mathbf{n}]\}$ is a finite discrete group and we can discuss the irreducible matrix representations in terms of the theory developed in Part 1. The irreducible representations of space groups are, however, very peculiar entities and require detailed exposition.

We introduce the irreducible representations of space groups by a discussion of the necessary and sufficient conditions that a vector X shall be such that the set of vectors $\{[\alpha|\mathbf{v}(\alpha)+\mathbf{n}]X\}$ spans a space S, irreducible under the space group $G \equiv \{[\alpha|\mathbf{v}(\alpha)+\mathbf{n}]\}$. The nature of the space S need not concern us here, beyond the requirement that the multiplication rule (92) be valid in S. These conditions were first discovered by Seitz (1936). His treatment is so obscure, however, that it is unsuitable as a reference here. Moreover Seitz does not consider the double

space groups, and his treatment appears to be limited to simple space groups, that is groups that do not contain glide-plane and screw symmetry operators.

We base our discussion on the coset theorems given in the Appendix. To this end, we consider first the irreducible representation of the sub-group of translations $T \equiv \{[\epsilon|\mathbf{n}]\}$, and to do this, it is convenient to introduce the lattice L^{-1}, reciprocal to the lattice L defining T.

The reciprocal lattice.

Given a crystal lattice L, basis $\{a_r\}$, point group (α), we define the lattice L^{-1}, reciprocal to L, as the lattice with basis $\{b_r\}$, where $b_r = [a_s \times a_t]/a_r.[a_s \times a_t]$. With this definition, we have at once the orthogonality relations $b_r.a_s = \delta_{rs}$. Now let $\mathbf{K} = \Sigma_r b_r K_r$ denote a vector in L^{-1}, that is, K_r is integral, then

$$\mathbf{K}.\mathbf{n} = \sum_{r,s}(b_r.a_s)K_r n_s = \sum_r K_r n_r. \qquad \ldots\ldots(94)$$

It follows that $\exp 2\pi i(\mathbf{K}.\mathbf{n}) = 1$ for all $\mathbf{K} \in L^{-1}$ and $\mathbf{n} \in L$.

Given the vector $\mathbf{K} \in L^{-1}$, we can show that $\alpha\mathbf{K} \in L^{-1}$, by calculating the coordinates of $\alpha\mathbf{K}$ with respect to $\{b_r\}$; for let $\alpha\mathbf{K} = \Sigma_s b_s k_s$, then

$$k_r = \alpha\mathbf{K}.a_r = \mathbf{K}.\alpha^{-1}a_r = \text{integer } N_r, \text{ from (94), since } \begin{cases}\mathbf{K} \in L^{-1} \\ \alpha^{-1}a_r \in L.\end{cases}$$

It follows that the point group of L^{-1} is identical with the point group of L.

Representations of the translation sub-group T.

Consider the set of complex numbers $\{\exp 2\pi i(\mathbf{k}.\mathbf{n})\}$ corresponding to a given vector $\mathbf{k} = \Sigma_r b_r k_r$. These numbers afford a representation of T, for

$$\exp 2\pi i(\mathbf{k}.\mathbf{m}) \exp 2\pi i(\mathbf{k}.\mathbf{n}) = \exp 2\pi i(\mathbf{k}.\mathbf{m}+\mathbf{n}).$$

In order to satisfy the modulo condition (93), we impose the constraint that the coordinates k_r of $\mathbf{k}$ shall be rational numbers

$$k_r = \frac{q_r}{\nu_r}, \qquad \ldots\ldots(95)$$

where the integers ν_r define the super-lattice vectors $\mathbf{N}$ of (93); that is

$$\mathbf{N} = \sum_r p_r(\nu_r a_r), \qquad \ldots\ldots(96)$$

where the coordinates p_r are integral. We have then,

$$\exp 2\pi i(\mathbf{k}.\mathbf{n}) = \exp 2\pi i(\mathbf{k}.\mathbf{m}) \quad \text{if } \mathbf{n}-\mathbf{m} = \mathbf{N}, \qquad \ldots\ldots(97)$$

in agreement with (93). Since there are $\nu_1\nu_2\nu_3$ crystal lattice points within a unit cell of the super-lattice, it follows that the order of the group T under the modulo condition is $\nu_1\nu_2\nu_3$. Furthermore, the number of classes of equivalent element in T under the modulo condition, and therefore the number of irreducible characters, is $\nu_1\nu_2\nu_3$, for each element is in a class by itself, since from (92)

$$[\epsilon|\mathbf{n}]^{-1}[\epsilon|\mathbf{m}][\epsilon|\mathbf{n}] = [\epsilon|\mathbf{m}].$$

A group of this form is sometimes called an Abelian group. The group T is in fact a cyclic Abelian group, since

$$[\epsilon|\,\mathbf{n}] = [\epsilon|\,a_1 n_1 + a_2 n_2 + a_3 n_3] = [\epsilon|\,a_1]^{n_1}\,[\epsilon|\,a_2]^{n_2}\,[a_3]^{n_3}.$$

The constraint (95) on $\mathbf{k}$ effectively defines a sub-lattice $\{\mathbf{k}\}$ of the reciprocal lattice L^{-1}. Moreover, since

$$\exp 2\pi i(\mathbf{k}.\mathbf{n}) = \exp 2\pi i(\mathbf{k}+\mathbf{K}.\mathbf{n}) \qquad \ldots\ldots(98)$$

from (94), we have the modulo condition on $\mathbf{k}$,

$$\mathbf{k}' \text{ and } \mathbf{k}'' \text{ are equivalent if } \mathbf{k}'-\mathbf{k}'' = \mathbf{K}. \qquad \ldots\ldots(99)$$

To each distinct value of $\mathbf{k}$ there corresponds a representation $\{\exp 2\pi i(\mathbf{k}.\mathbf{n})$ of T. These $\nu_1\nu_2\nu_3$ representations form a complete set of irreducible representations of T, for they satisfy the irreducible character relations

$$\sum_{\mathbf{n}} \exp 2\pi i(\mathbf{k}.\mathbf{n}) \exp[-2\pi i(\mathbf{k}'.\mathbf{n})] = \delta_{\mathbf{k}\mathbf{k}'}\quad \nu_1\nu_2\nu_3\,; \qquad \ldots\ldots(100)$$

that is, all the irreducible representations of T are one-dimensional. This is a general property of Abelian groups; for we know from our discussion of the decomposition of the regular representation of a finite group, that $\Sigma_i n_i^2 = N$, and the number of terms in the sum Σ_i is N for an Abelian group, so $n_i = 1$ for all i, since n_i is real.

The vectors $X^{\mathbf{k}}$.

In order to denote that a vector X affords the irreducible representation $\{\exp 2\pi i(\mathbf{k}.\mathbf{n})\}$ of T, we shall attach a superscript and write $X^{\mathbf{k}}$; then by definition,

$$[\epsilon|\,\mathbf{n}]\,X^{\mathbf{k}} = X^{\mathbf{k}} \exp 2\pi i(\mathbf{k}.\mathbf{n}). \qquad \ldots\ldots(101)$$

The vectors $X^{\mathbf{k}}$ possess the property that $[\alpha|\,\mathbf{v}(\alpha)]\,X^{\mathbf{k}}$ is of the form $X^{\alpha\mathbf{k}}$; for from (92),

$$\left.\begin{aligned}[\epsilon|\,\mathbf{n}]\,[\alpha|\,\mathbf{v}(\alpha)]\,X^{\mathbf{k}} &= [\alpha|\,\mathbf{v}(\alpha)]\,[\epsilon|\,\alpha^{-1}\,\mathbf{n}]\,X^{\mathbf{k}},\\ &= [\alpha|\,\mathbf{v}(\alpha)]\,X^{\mathbf{k}} \exp 2\pi i(\mathbf{k}.\alpha^{-1}\,\mathbf{n}),\\ &= [\alpha|\,\mathbf{v}(\alpha)]\,X^{\mathbf{k}} \exp 2\pi i(\alpha\mathbf{k}.\mathbf{n}),\end{aligned}\right\} \qquad \ldots\ldots(102)$$

since α is unitary. It is important to note here, that in forming the vector $\alpha\mathbf{k}$ from $\mathbf{k}$, we operate with the geometrical operator α, taken from the geometrical point group, *not* from the double point group. If α' and α'' are elements of the double point group, with the same image in the geometrical point group, then

$$[\alpha'|\,\mathbf{v}(\alpha)]\,X^{\mathbf{k}} = Y^{\mathbf{k}}\,[\alpha''|\,\mathbf{v}(\alpha)]\,X^{\mathbf{k}}Z^{\mathbf{k}}. \qquad \ldots\ldots(103)$$

Spaces irreducible under a space group.

We can now take up our main discussion again. We consider first the case of the substitutional space groups, and begin with the assertion that the space spanned by the set of vectors $\{[\alpha|\,\mathbf{v}(\alpha)]\,X^{\mathbf{k}}\}$ is irreducible under the space group $\{[\alpha|\,\mathbf{v}(\alpha)+\mathbf{n}]\}$, if the vector $\mathbf{k}$ is such that $\alpha\mathbf{k}$ and $\beta\mathbf{k}$ are *not* equivalent vectors, for all $\alpha \neq \beta$ in the point group; that is

$$\alpha\mathbf{k} \neq \beta\mathbf{k}+\mathbf{K}, \quad \text{if} \quad \alpha \neq \beta. \qquad \ldots\ldots(104)$$

This is a sufficient condition. It is not satisfied in the case of the double space groups, since α and β may be elements of the double point group, with the same image in the geometrical point group, so that $\alpha\mathbf{k} = \beta\mathbf{k}, \alpha \neq \beta$. When we consider the case of double space groups we shall relax (104). In order to prove our assertion of irreducibility, we invoke the theory of cosets.

Assuming that the reader has studied the theorems on cosets in the Appendix, we observe that the space group operations of the form $[\alpha|\mathbf{v}(\alpha)]$ can be chosen as the left-coset representatives of the group $\{[\alpha|\mathbf{v}(\alpha)+\mathbf{n}]\}$, with respect to the sub-group of translations T; that is, we identify T with the sub-group H of the theorems, and $[\alpha|\mathbf{v}(\alpha)]$ with the left-coset representative b_α. We identify the sub-space S_α of the theorems with the one-dimensional space spanned by the vector $[\alpha|\mathbf{v}(\alpha)]X^{\mathbf{k}}$. Two such spaces, S_α and $S_\beta, \alpha \neq \beta$, are mutually orthogonal, for they afford inequivalent irreducible representation of T, since, by hypothesis, $\alpha\mathbf{k} \neq \beta\mathbf{k} + \mathbf{K}$. The irreducibility condition

$$\sum_{\xi} \chi_\alpha(\xi)\chi^*{}_\beta(\xi) = 0, \quad \alpha \neq \beta$$

is satisfied, for the group of elements $H_{\alpha\beta} \equiv \{\xi\}$, common to H_α and H_β, is $H \equiv T$ itself, since

$$[\alpha|\mathbf{v}(\alpha)]\,[\epsilon|\mathbf{n}]\,[\alpha|\mathbf{v}(\alpha)]^{-1} = [\epsilon|\alpha\mathbf{n}]\,;$$

that is, $T = H_\alpha = H_\beta = \ldots$; and χ_α and χ_β are inequivalent irreducible characters of $H_{\alpha\beta} \equiv T$.

For general values of $\mathbf{k}$ then, that is, values of $\mathbf{k}$ satisfying (104), it follows that the irreducible representation of the substitutional space groups are completely characterized by the chosen value of $\mathbf{k}$ labelling the initial vector $X^{\mathbf{k}}$. The basis of a given irreducible representation of a space group $\{[\alpha|\mathbf{v}(\alpha)+\mathbf{n}]\}$ is the set of basis vectors $\{[\alpha|\mathbf{v}(\alpha)]X^{\mathbf{k}}\}$. It is evident that we can take any one of these basis vectors as our initial vector, so the irreducible representation can be characterized by any one of the set of $\mathbf{k}$-vectors $\{\alpha\mathbf{k}\}$. In order to construct a basis vector of the form $u(p,r,i)$ discussed in Part 1, §§ 6 and 7, relative to the space group $\{[\alpha|\mathbf{v}(\alpha)+\mathbf{n}]\}$, it is sufficient to construct a vector of the form $X^{\mathbf{k}}$ satisfying (101), where $\mathbf{k}$ satisfies (104). This we can do by using the corresponding projection operator

$$\epsilon^{\mathbf{k}} = \frac{1}{\nu_1\nu_2\nu_3}\sum_{\mathbf{n}}[\epsilon|\mathbf{n}]\exp -2\pi i(\mathbf{k}.\mathbf{n}). \qquad \ldots\ldots(105)$$

Before we consider irreducible representations of space groups, corresponding to special values of $\mathbf{k}$, we generalize the foregoing argument to the double space groups.

Extension to the double space groups.

In the case of the double space groups, we know from Part 2, § 2, that the coset-representatives will occur in pairs, say $[\alpha'|\mathbf{v}(\alpha)]$ and $[\alpha''|\mathbf{v}(\alpha)]$, where α' and α'' map on the same image in the geometrical point group. We now modify the theorems on cosets in the Appendix by taking the space S to be a sum of the sub-spaces S_α, such that either $S_{\alpha'}$ or $S_{\alpha''}$ is included in the sum, but not both; the reason for this exclusion being that under the unimodular restriction

$S_{\alpha'}$ and $S_{\alpha''}$ are identical. When we re-examine the condition that S shall be irreducible under G, we can see that the term $\sum_{\alpha+\beta}\sum_g \chi_\alpha(g)\chi^*{}_\beta(g)$ is zero, for the same reason as before. Now consider the term

$$\sum_\alpha \sum_g \chi_\alpha(g)\chi^*{}_\alpha(g).$$

The sum Σ_α is now to be taken over only half the coset-representatives. We gain a factor 2 in the sum Σ_g, however, since $b_{\alpha'}u_q = -b_{\alpha''}u_q$, where the negative sign is absorbed in $|\chi_\alpha(g)|^2$. It follows that the conditions for irreducibility are the same for the double space groups as for the substitutional space groups, provided we interpret α and β in condition (104) as elements of the geometrical point group.

Special values of K.

The discussion for special values of $\mathbf{k}$ is a generalization of the foregoing argument, the essential point being that we define the group to be chosen as the sub-group $H = H_1$ of the coset theorems, in a way that includes the choice $H = T$ as a special case. Consider then, the set of point group operators γ defined by the relation

$$\gamma\mathbf{k} = \mathbf{k} + \mathbf{K}, \qquad \ldots\ldots(106)$$

for a given value of $\mathbf{k}$. The set $\{\gamma\}$ is a sub-group of the point group, for

$$\gamma \in \{\gamma\} \quad \text{implies } \gamma^{-1} \in \{\gamma\}, \quad \text{since} \quad \gamma\mathbf{k} = \mathbf{k} + \mathbf{K} \quad \text{implies} \quad \gamma^{-1}\mathbf{k} = \mathbf{k} + \mathbf{K}';$$

$$\left.\begin{matrix}\gamma \in \{\gamma\} \\ \delta \in \{\gamma\}\end{matrix}\right\} \text{ implies } \begin{matrix}\gamma\delta \in \{\gamma\} \\ \delta\gamma \{\in \gamma\}\end{matrix} \text{ for } \begin{matrix}\gamma\delta\mathbf{k} = \mathbf{k} + \mathbf{K}'', \\ \delta\gamma\mathbf{k} = \mathbf{k} + \mathbf{K}''.\end{matrix}$$

It follows that a general value of $\mathbf{k}$ defined by (104), can be defined by the equivalent condition that the sub-group $\{\gamma\}$, corresponding to a general value of $\mathbf{k}$, contains only the unit element. Since $\{\gamma\}$ is a sub-group of the point group of the space group, the set of elements $\{[\gamma|\mathbf{v}(\gamma)+\mathbf{n}]\}$ is a sub-group of the complete space group. We propose to take this sub-group as our sub-group $H = H_1$. Let $\{[\alpha|\mathbf{v}(\alpha)]\}$ now denote the set of left-coset representatives of the complete space group relative to this sub-group $\{[\gamma|\mathbf{v}(\alpha)+\mathbf{n}]\}$ where, in the case of the double space groups, we impose the same restriction as before; namely, that $\{[\alpha|\mathbf{v}(\alpha)]\}$ shall contain only one member of each pair $[\alpha'|\mathbf{v}(\alpha)]$ and $[\alpha''|\mathbf{v}(\alpha)]$, if α' and α'' map on the same element in the geometrical point group.

In order to satisfy the conditions imposed in the coset theorems, we choose the sub-space S_1 to be irreducible under the group $\{[\gamma|\mathbf{v}(\gamma)+\mathbf{n}]\}$ defined by (106). Moreover, we impose the condition that S_1 shall contain at least one vector of the form $X^{\mathbf{k}}$, corresponding to the given value of $\mathbf{k}$. It then follows from (102) and (106), that every vector in S_1 is of the form $X^{\mathbf{k}}$, for we can generate S_1 from a vector of the form $X^{\mathbf{k}}$, since S_1 is irreducible under $\{[\gamma|\mathbf{v}(\gamma)+\mathbf{n}]\}$. Furthermore, this construction implies that every vector in the sub-space S_α is of the form $X^{\alpha\mathbf{k}}$. We can now prove that the sub-spaces $S_\alpha, S_\beta, \ldots$ are mutually orthogonal. Since every vector in S_α is of the form $X^{\alpha\mathbf{k}}$ and every vector in S_β is of the form $X^{\beta\mathbf{k}}$, the mutual orthogonality of S_α and S_β follows if $\alpha\mathbf{k} \neq \beta\mathbf{k} + \mathbf{K}$, for then $X^{\alpha\mathbf{k}}$ and $X^{\beta\mathbf{k}}$ afford inequivalent irreducible representations of the translation sub-group T.

The required relation $\alpha\mathbf{k} \neq \beta\mathbf{k} + \mathbf{K}$ follows from the definition of the coset-representatives $[\alpha|\mathbf{v}(\alpha)]$ and $[\beta|\mathbf{v}(\beta)]$; for $\alpha\mathbf{k} = \beta\mathbf{k} + \mathbf{K}$ implies $\beta^{-1}\alpha\mathbf{k} = \mathbf{k} + \mathbf{K}'$, that is $\beta^{-1}\alpha \in \{\gamma\}$, i.e. $\alpha = \beta\gamma$, contrary to the hypothesis that $[\alpha|\mathbf{v}(\alpha)]$ and $[\beta|\mathbf{v}(\beta)]$ are distinct coset-representatives.

In order to complete the proof of the irreducibility of the space $S = \Sigma_\alpha S_\alpha$ under the complete space group, we have to show that the sum $\Sigma_\xi \chi_\alpha(\xi)\chi^*_\beta(\xi)$, appearing in the coset-theorems, is zero if $\alpha \neq \beta$. In order to prove this, we observe that in the decomposition of the possibly reducible characters χ_α and χ_β of $H_{\alpha\beta}$, no irreducible character of $H_{\alpha\beta}$ will be found to be common to χ_α and χ_β, for $H_{\alpha\beta}$ contains T as a sub-group, and no irreducible representation of T is common to S_α and S_β. It follows then from the character orthogonality relations, that $\Sigma_\xi \chi_\alpha(\xi)\chi^*_\beta(\xi) = 0$ if $\alpha \neq \beta$.

Since the foregoing discussion is applicable to all values of $\mathbf{k}$, we can, if we wish, dispense with the discussion based on (104).

In order to show that all irreducible representations of space groups are accessible by this construction, we observe that given any space S irreducible under a space group, we can find a vector in S of the form $X^{\mathbf{k}}$ for some value of $\mathbf{k}$, such that the set of vectors $\{[\gamma|\mathbf{v}(\gamma)]X^{\mathbf{k}}\}$ corresponding to $\mathbf{k}$, spans a space $S_{\mathbf{k}}$ irreducible under the corresponding sub-group $\{[\gamma|\mathbf{v}(\gamma)+\mathbf{n}]\}$. By means of the given procedure, we can now construct a sub-space S′ of S, irreducible under the given space group. But S is irreducible by hypothesis, therefore S′ is identical to S.

We can summarize our discussion in the assertion that the necessary and sufficient condition that a vector X shall be such that the set of vectors $\{[\alpha|\mathbf{v}(\alpha)+\mathbf{n}]X\}$ spans a space irreducible under the space group $\{[\alpha|\mathbf{v}(\alpha)+\mathbf{n}]\}$, is that X shall be of the form $X^{\mathbf{k}}$, where the set of vectors $\{[\gamma|\mathbf{v}(\gamma)+\mathbf{n}]X^{\mathbf{k}}\}$ spans a space irreducible under the sub-group $\{[\gamma|\mathbf{v}(\gamma)+\mathbf{n}]\}$ defined by the condition $\gamma\mathbf{k} = \mathbf{k} + \mathbf{K}$.

It is evident that this conclusion is of practical value only if it is possible to find the irreducible representation of the sub-group $\{[\gamma|\mathbf{v}(\gamma)+\mathbf{n}]\}$. We have shown that this is a trivial problem in the case of general values of $\mathbf{k}$, for then $\{[\gamma|\mathbf{v}(\gamma)+\mathbf{n}]\} \equiv T$. We now consider certain curious properties of the matrix representations of the group $\{[\gamma|\mathbf{v}(\gamma)+\mathbf{n}]\}$, which simplify the problem in the case of special values of $\mathbf{k}$.

Representations of the group $\{[\gamma|\mathbf{v}(\gamma)+\mathbf{n}]\}$.

We have imposed the condition that the sub-space S_1 of the foregoing discussion shall be irreducible under the sub-group $\{[\gamma|\mathbf{v}(\gamma)+\mathbf{n}]\}$, with the further condition that every vector in S_1 is of the form $X^{\mathbf{k}}$, where $\mathbf{k}$ and $\{\gamma\}$ are related by the condition $\gamma\mathbf{k} = \mathbf{k} + \mathbf{K}$. It follows from the latter condition that in a matrix representation afforded by the space S_1, the matrix representative of an element $[\gamma|\mathbf{v}(\gamma)]$ commutes with the representative of a translation element $[\epsilon|\mathbf{n}]$; for

$$[\gamma|\mathbf{v}(\gamma)]\,[\epsilon|\mathbf{n}]\,X^{\mathbf{k}} = [\gamma|\mathbf{v}(\gamma)]\,X^{\mathbf{k}} \exp 2\pi i(\mathbf{k}.\mathbf{n}),$$

and

$$[\epsilon|\mathbf{n}]\,[\gamma|\mathbf{v}(\gamma)]\,X^{\mathbf{k}} = [\gamma|\mathbf{v}(\gamma)]\,X^{\mathbf{k}} \exp 2\pi i(\gamma\mathbf{k}.\mathbf{n}), \quad \text{from (102)};$$

but $\gamma\mathbf{k} = \mathbf{k} + \mathbf{K}$, by hypothesis, so

$$[\gamma|\mathbf{v}(\gamma)]\,[\epsilon|\mathbf{n}]\,X^{\mathbf{k}} = [\epsilon|\mathbf{n}]\,[\gamma|\mathbf{v}(\gamma)]\,X^{\mathbf{k}}. \qquad \ldots\ldots(107)$$

The group $G_{\mathbf{k}}$.

This commutation relation (107) implies that we can replace the condition that the space S_1 be irreducible under the group $\{[\gamma|\mathbf{v}(\gamma)+\mathbf{n}]\}$, by the condition that S_1 be irreducible under the sub-group $G_{\mathbf{k}}$ generated by the elements $\{[\gamma|\mathbf{v}(\gamma)]\}$. If the vectors $\mathbf{v}(\gamma)$ are not all zero, this sub-group contains certain elements of T, but in practice all but a few of these translation elements possess the same matrix representatives, because only simple values of $\mathbf{k}$ satisfy $\gamma\mathbf{k}=\mathbf{k}+\mathbf{K}$. In fact the group $G_{\mathbf{k}}$ is generally isomorphous with one of the crystal point groups. Given an irreducible representation of $G_{\mathbf{k}}$, we can construct the corresponding projection operators and then apply these to a vector of the form $X^{\mathbf{k}}$, where the vector $\mathbf{k}$ satisfies $\gamma\mathbf{k}=\mathbf{k}+\mathbf{K}$ for all γ. By this means we can construct a basis in the space S_1. Not all the irreducible representations of the group $G_{\mathbf{k}}$ are acceptable, however, for the constraints already imposed limit our choice to those representations in which the pure translation elements, say $[\epsilon|\mathbf{m}]$, of $G_{\mathbf{k}}$ are represented by the unit matrix multiplied by $\exp 2\pi i(\mathbf{k}.\mathbf{m})$.

The problem is especially simple if all the vectors $\mathbf{v}(\gamma)$ are zero, for then $G_{\mathbf{k}}$ is isomorphous with the group $\{\gamma\}$. In this case, the irreducible representation of the complete space group $\{[\alpha|\mathbf{v}(\alpha)+\mathbf{n}]\}$ can be characterized by $\mathbf{k}$ and the irreducible characters of the corresponding group $\{\gamma\}$. This simple result was obtained by Bouckaert, Smoluchowski and Wigner (1936).

Energy bands.

Given a Hamiltonian with space-group symmetry, it is convenient to associate the eigenvalues and eigenvectors with points in a $\mathbf{k}$-space containing a reference lattice isomorphous with the lattice L^{-1}, reciprocal to the lattice of the given space group. We denote the eigenvalues by $E(\mathbf{k})$, where $E(\mathbf{k})$ is a multi-valued quasi-continuous function of $\mathbf{k}$, since the values of $\mathbf{k}$ allowed by the cyclic boundary conditions can be taken as close as we please, by increasing the size of the super-lattice unit cell. This distribution of energy levels forms the energy bands of the system.

The function $E(\mathbf{k})$ has the symmetry of the space group $\{[\alpha|\mathbf{K}]\}$ based on the reciprocal lattice L^{-1}; for it follows from the structure of the irreducible representation of the space group of the Hamiltonian, that $E(\mathbf{k})=E(\mathbf{k}+\mathbf{K})$ and $E(\alpha\mathbf{k})=E(\mathbf{k})$. For special values of $\mathbf{k}$, the structure of the degeneracy will become more complex if a given eigenvalue $E(\mathbf{k})$ corresponds to a multi-dimensional irreducible representation of $G_{\mathbf{k}}$, for in this case, two or more energy bands 'touch' at the point $\mathbf{k}$; and at corresponding points $\alpha\mathbf{k}$, determined by the coset-representatives $[\alpha|\mathbf{v}(\alpha)]$. If we interpret the condition $\gamma\mathbf{k}=\mathbf{k}+\mathbf{K}$ geometrically, we can regard these special values of $\mathbf{k}$ as points on symmetry planes and axes defined by the reciprocal space group $\{[\alpha|\mathbf{K}]\}$. Further complexity may be introduced by time-reversal symmetry which we now consider.

Time-reversal symmetry.

Wigner's time-reversal theorems were applied to the substitutional space groups by Herring (1937) and extended to the double space groups by Elliott (1954). We can treat both kinds of space group together, provided we make the

same reservations over coset-representatives as we did in the discussion of the irreducible representations of the double space groups.

In order to evaluate Wigner's sum $\Sigma_P \chi^i(P^2)$ for the general space group $\{[\alpha|\,\mathbf{v}(\alpha)+\mathbf{n}]\}$, we observe that

$$[\alpha|\,\mathbf{v}(\alpha)+\mathbf{n}]^2 = [\epsilon|\,\mathbf{n}+\alpha\mathbf{n}]\,[\alpha|\,\mathbf{v}(\alpha)]^2, \qquad \ldots\ldots(108)$$

so the spur of the element $[\alpha|\,\mathbf{v}(\alpha)+\mathbf{n}]^2$ in a given irreducible representation of the space group $\{[\alpha|\,\mathbf{v}(\alpha)+\mathbf{n}]\}$, can be written

$$\sum_{\beta}\sum_{r}[\beta|\,\mathbf{v}(\beta)]\,X_r^{\mathbf{k}}.[\epsilon|\,\mathbf{n}+\alpha\mathbf{n}]\,[\alpha|\,\mathbf{v}(\alpha)]^2\,[\beta|\,\mathbf{v}(\beta)]\,X_r^{\mathbf{k}}, \qquad \ldots\ldots(109)$$

where $\{[\beta|\,\mathbf{v}(\beta)]\,X_r^{\mathbf{k}}\}$ is the basis chosen in the sub-space S_β, corresponding to the chosen irreducible representation. Now

$$\begin{aligned}&[\beta|\,\mathbf{v}(\beta)]\,X_r^{\mathbf{k}}.[\epsilon|\,\mathbf{n}+\alpha\mathbf{n}]\,[\alpha|\,\mathbf{v}(\alpha)]^2[\beta|\,\mathbf{v}(\beta)]\,X_r^{\mathbf{k}}\\ &\quad=[\epsilon|\,\mathbf{n}+\alpha\mathbf{n}]^{-1}[\beta|\,\mathbf{v}(\beta)]\,X_r^{\mathbf{k}}.[\alpha|\,\mathbf{v}(\alpha)]^2[\beta|\,\mathbf{v}(\beta)]\,X_r^{\mathbf{k}}, \qquad \ldots\ldots(110)\end{aligned}$$

since the operator $[\epsilon|\,\mathbf{n}+\alpha\mathbf{n}]$ is unitary. Furthermore, from the multiplication rule (92), and equation (102),

$$\begin{aligned}[\epsilon|\,\mathbf{n}+\alpha\mathbf{n}]^{-1}[\beta|\,\mathbf{v}(\beta)]\,X_r^{\mathbf{k}} &= [\beta|\,\mathbf{v}(\beta)]\,X_r^{\mathbf{k}}\exp\left[-2\pi i(\mathbf{k}.\beta^{-1}\,\mathbf{n}+\beta^{-1}\,\alpha\mathbf{n})\right],\\ &=[\beta|\,\mathbf{v}(\beta)]\,X_r^{\mathbf{k}}\exp\left[-2\pi i(\beta\mathbf{k}+\alpha^{-1}\beta\mathbf{k}.\,\mathbf{n})\right];\end{aligned}$$

for

$$\begin{aligned}(\mathbf{k}.\beta^{-1}\,\mathbf{n}+\beta^{-1}\,\alpha\mathbf{n}) &= (\mathbf{k}.\beta^{-1}\,\mathbf{n})+(\mathbf{k}.\beta^{-1}\,\alpha\mathbf{n})\\ &=(\beta\mathbf{k}.\,\mathbf{n})+(\alpha^{-1}\beta\mathbf{k}.\,\mathbf{n}).\end{aligned}$$

It follows that the spur of $[\alpha|\,\mathbf{v}(\alpha)+\mathbf{n}]^2$ is

$$\sum_{\beta}\sum_{r}[\beta|\,\mathbf{v}(\beta)]\,X_r^{\mathbf{k}}.[\alpha|\,\mathbf{v}(\alpha)]^2[\beta|\,\mathbf{v}(\beta)]\,X_r^{\mathbf{k}}\exp 2\pi i(\beta\mathbf{k}+\alpha^{-1}\beta\mathbf{k}.\,\mathbf{n}). \qquad \ldots\ldots(111)$$

The sum over the elements of the space group can be decomposed into a sum over the elements of T, and a sum over the coset-representatives $[\alpha|\,\mathbf{v}(\alpha)]$ with respect to T. Thus Wigner's sum can be written

$$\sum_{\alpha,\beta}\sum_{r}[\beta|\,\mathbf{v}(\beta)]\,X_r^{\mathbf{k}}.[\alpha|\,\mathbf{v}(\alpha)]^2[\beta|\,\mathbf{v}(\beta)]\,X_r^{\mathbf{k}}\sum_{n}\exp 2\pi i(\beta\mathbf{k}+\alpha^{-1}\beta\mathbf{k}.\,\mathbf{n}), \qquad \ldots\ldots(112)$$

and we know from the time-reversal theorems that this sum is either zero or equal to $\pm N$, where N is the order of the space group under the modulo condition. The order N is equal to $\nu_1\nu_2\nu_3$ times the order of the point group. We observe that

$$\sum_{n}\exp 2\pi i(\beta\mathbf{k}+\alpha^{-1}\beta\mathbf{k}.\,\mathbf{n})=\nu_1\nu_2\nu_3$$

if

$$\beta\mathbf{k}+\alpha^{-1}\beta\mathbf{k}=\mathbf{K}, \qquad \ldots\ldots(113)$$

and is otherwise zero. Furthermore, from (102),

$$[\beta|\,\mathbf{v}(\beta)]\,X_r^{\mathbf{k}}.[\alpha|\,\mathbf{v}(\alpha)]^2[\beta|\,\mathbf{v}(\beta)]\,X_r^{\mathbf{k}}$$

is zero unless

$$\beta^{-1}\alpha^2\beta\mathbf{k}=\mathbf{k}+\mathbf{K}'. \qquad \ldots\ldots(114)$$

However, (114) is satisfied if α satisfies (113); for (113) implies

$$\beta^{-1}\alpha^2\beta\mathbf{k}+\beta^{-1}\alpha\beta\mathbf{k}=\mathbf{K}'', \qquad \ldots\ldots(115)$$

and

$$\beta^{-1}\alpha\beta\mathbf{k}+\mathbf{k}=\mathbf{K}''. \qquad \ldots\ldots(116)$$

So if we subtract (116) from (115) we obtain (114) in the form

$$\beta^{-1}\alpha^2\beta\mathbf{k}-\mathbf{k}=\mathbf{K}^{IV}. \qquad \ldots\ldots(117)$$

As a further simplification, we note that if $\alpha=\gamma$ is a solution to $\alpha\mathbf{k}+\mathbf{k}=\mathbf{K}$, then $\alpha=\beta\gamma\beta^{-1}$ is a solution to $\alpha\beta\mathbf{k}+\beta\mathbf{k}=\mathbf{K}'$.

In the case of the double space groups we know that only half the abstract coset-representatives $[\beta|\mathbf{v}(\beta)]$ appear in the sum Σ_β in (112), but we obtain a factor 2 in the sum Σ_α, because two elements of the double group correspond to each solution to the equation $\gamma\mathbf{k}+\mathbf{k}=\mathbf{K}$. (The negative sign is absorbed in forming the square.)

For general values of $\mathbf{k}$ it is evident that the only solution to $\gamma\mathbf{k}+\mathbf{k}=\mathbf{K}$ would be $\gamma=J$, the inversion, for $J\mathbf{k}=-\mathbf{k}$. It follows, if J is not an element of the point group, that Wigner degeneracy occurs, corresponding to the case of a complex character. We can see that in this case, the time-reversal operator $\mathscr{R}$ induces degeneracy between states labelled with $\mathbf{k}$ and $-\mathbf{k}$, for $\mathscr{R}$ commutes with $[\epsilon|\mathbf{n}]$ from (40), so

$$[\epsilon|\mathbf{n}]\mathscr{R}X^{\mathbf{k}}=\exp[-2\pi i(\mathbf{k}.\mathbf{n})]\mathscr{R}X^{\mathbf{k}}. \qquad \ldots\ldots(118)$$

If the inversion J is an element of the point group, then $[J|\mathbf{v}(J)]^2=[\epsilon|0]$, and the Wigner sum is equal to N. It follows that in this case Wigner degeneracy does not occur if the number of electrons in the system is even, or we neglect spin–orbit coupling. If the number of electrons is odd and we include spin–orbit coupling, Wigner degeneracy occurs, corresponding to the case $(\chi^i)^*=\chi^i$ and $\mathscr{R}S^i\neq S^i$ of (44). This is the most remarkable deduction from Wigner's theorems, for without his analysis, such degeneracy could go unsuspected.

For special values of $\mathbf{k}$ the discussion is simple if the vectors $\mathbf{v}(\gamma)$ in the elements $[\gamma|\mathbf{v}(\gamma)]$ satisfying $\gamma\mathbf{k}+\mathbf{k}=\mathbf{K}$ are all zero, for the Wigner sum is then equal to $+N$, and degeneracy occurs only if the number of electrons is odd and spin–orbit coupling is included. If the vectors $\mathbf{v}(\gamma)$ are not zero, as in the case of a screw-diad symmetry element, we must be cautious, for we observe that in (112)

$$\begin{aligned}&[\beta|\mathbf{v}(\beta)]X_r^{\mathbf{k}}.[\alpha|\mathbf{v}(\alpha)]^2[\beta|\mathbf{v}(\beta)]X_r^{\mathbf{k}}\\&\quad=X_r^{\mathbf{k}}.[\beta|\mathbf{v}(\beta)]^{-1}[\alpha|\mathbf{v}(\alpha)]^2[\beta|\mathbf{v}(\beta)]X_r^{\mathbf{k}}, \qquad \ldots\ldots(119)\end{aligned}$$

so if, for example, $[\alpha|\mathbf{v}(\alpha)]=[\gamma|\mathbf{v}(\gamma)]$ represents a screw-diad, then

$$[\gamma|\mathbf{v}(\gamma)]^2=[\epsilon|\mathbf{m}],$$

and we obtain a factor $\exp 2\pi i(\mathbf{k}.\mathbf{m})$ in the evaluation of (119). This factor will be equal to -1 if $\mathbf{k}$ and $\mathbf{m}$ are such that $(\mathbf{k}.\mathbf{m})$ is half integral. This factor must be the same for all β and $\alpha=\gamma$ in the sum (112), if (112) is non-zero, for the sum itself is then exactly equal to $\pm N$. A detailed analysis of various examples is given by Elliott (1954).

Acknowledgments

The author would like to thank Professor B. H. Flowers, Drs. J. P. Elliott, W. M. Lomer, R. D. Lowde and F. Mandl for detailed criticism of various drafts of Part 1 of the article. Part 2 has benefited from discussions with Mr. John Gabriel.

APPENDIX I

Proof of the generalized orthogonality relations.

There are four parts to the proof of the relations

$$\sum_A U_{rs}^i(A)\, U_{\alpha\beta}^j(A^{-1}) = \frac{N}{n_i}\,\delta_{r\beta}\,\delta_{s\alpha}\,\delta_{ij} \qquad \ldots\ldots\text{(A1)}$$

introduced in Part 1, §6, corresponding to the three Kronecker symbols, and the case when all three Kronecker conditions are simultaneously satisfied. We leave till last the case $i \neq j$, since this involves a generalization of the form of Schur's lemma proved in Part 1, §4. The form of the lemma given there is sufficient to prove the first three parts of these relations. Consider then the sum

$$\sum_A U_{pq}^i(A)\, U_{rs}^i(A^{-1}),$$

where the set of matrices $\{U^i(A)\}$ affords an irreducible representation of the finite group $\{A\}$. Since the form of this sum is inconvenient, let us consider the $n_i \times n_i$ matrix

$$P = \sum_A U^i(A)\, Q\, U^i(A^{-1}), \qquad \ldots\ldots\text{(A2)}$$

where Q is an *arbitrary* $n_i \times n_i$ matrix. Our sums $\Sigma_A\, U_{pq}^i(A)\, U_{rs}^i(A^{-1})$ are now coefficients in linear combinations of the elements of the arbitrary matrix Q. The essence of Schur's proof is to demonstrate, by using his lemma, that the matrix P above has certain special properties. Since the elements of the matrix Q are arbitrary, these properties imply severe constraints on the coefficients

$$\sum_A U_{pq}^i(A)\, U_{rs}^i(A^{-1}),$$

relating the elements of P and Q.

We note first that the matrix P commutes with every matrix of the irreducible representation $\{U^i(A)\}$, for

$$\begin{aligned} U^i(B)P &= \sum_A U^i(B)\, U^i(A)\, Q\, U^i(A^{-1})\, U^i(B^{-1})\, U^i(B) \\ &= P\, U^i(B) \quad \text{since } \{A\} \text{ is a group.} \qquad \ldots\ldots\text{(A3)} \end{aligned}$$

It follows at once, from the matrix form of Schur's lemma, that

$$P = cI, \qquad \ldots\ldots\text{(A4)}$$

where I denotes the unit $n_i \times n_i$ matrix. In particular, the off-diagonal elements of P, namely

$$\sum_{q,r}\sum_A U_{pq}^i(A)\, Q_{qr}\, U_{rs}^i(A^{-1}), \quad p \neq s, \qquad \ldots\ldots\text{(A5)}$$

are zero, independent of the arbitrary elements of Q. It follows that

$$\sum_A U_{pq}^i(A)\, U_{rs}^i(A^{-1}) = 0 \quad \text{if } p \neq s, \text{ for all } q, r. \qquad \ldots\ldots\text{(A6)}$$

Furthermore, since $\{A\}$ is a group, this expression is unchanged if we replace A by A^{-1}; that is,

$$\sum_A U_{rs}^i(A)\, U_{pq}^i(A^{-1}) = 0 \quad \text{if } p \neq s, \text{ for all } q, r. \qquad \ldots\ldots\text{(A7)}$$

This result accounts for two of the Kronecker conditions in (A1).

Now let us consider the diagonal elements of (A4) namely

$$\sum_A \sum_{q,r} U_{pq}^i(A)\, Q_{qr}\, U_{rp}^i(A^{-1}) = c. \qquad \ldots\ldots(A8)$$

If we sum this relation over the suffix p, we obtain

$$N \sum_r Q_{rr} = n_i c, \qquad \ldots\ldots(A9)$$

where N is the order of the group $\{A\}$, for $\Sigma_p U_{pq}^i(A)\, U_{rp}^i(A^{-1}) = \delta_{qr}$. On the other hand, we know from (A7) that the only non-zero terms on the left-hand side of (A8) are those corresponding to $q = r$; that is,

$$\sum_A \sum_r U_{pr}^i(A)\, Q_{rr}\, U_{rp}^i(A^{-1}) = c. \qquad \ldots\ldots(A10)$$

Combining (A9) and (A10) to eliminate c,

$$\sum_r Q_{rr}\left[\frac{N}{n_i} - \sum_A U_{pr}^i(A)\, U_{rp}^i(A^{-1})\right] = 0. \qquad \ldots\ldots(A11)$$

But the elements Q_{rr} are arbitary, therefore

$$\sum_A U_{pr}^i(A)\, U_{rp}^i(A^{-1}) = \frac{N}{n_i}. \qquad \ldots\ldots(A12)$$

The fourth part of the proof of (A1) is the derivation of the relations

$$\sum_A U_{pq}^i(A)\, U_{rs}^j(A^{-1}) = 0, \quad \text{if } i \neq j, \text{ for all } p, q, r, s. \qquad \ldots\ldots(A13)$$

Again, we introduce an arbitrary matrix Q, but with the difference that here Q is an $n_i \times n_j$ matrix, so that the corresponding matrix

$$P = \sum_A U^i(A)\, Q\, U^j(A^{-1}) \qquad \ldots\ldots(A14)$$

is an $n_i \times n_j$ matrix. We now observe that

$$\begin{aligned} U^i(B)\, P &= \sum_A U^i(B)\, U^i(A)\, Q\, U^j(A^{-1})\, U^j(B^{-1})\, U^j(B), \\ &= P\, U^j(B), \qquad \ldots\ldots(A15) \end{aligned}$$

for every element B of the group $\{A\}$. The result (A13) then follows if we can prove that (A15) implies $P \equiv 0$, for the elements of Q in (A14) are arbitrary. This final step rests on the general form of Schur's lemma.

Schur's lemma.

We prove, if $\{U^i(A)\}$ and $\{U^j(A)\}$ are inequivalent irreducible representations of a set of operators $\{A\}$, that the relations

$$U^i(B)\, P = P\, U^j(B), \quad B \in \{A\}, \qquad \ldots\ldots(A16)$$

imply $P \equiv 0$.

Let the space S^i afford the given irreducible representation $\{U_{nm}^i(A)\}$ of $\{A\}$, relative to the basis $\{u_n\}$, $n = 1 \ldots n_i$; consider the sub-space of S^i spanned by the set of n_j vectors

$$\sum_n u_n P_{nr}.$$

This sub-space is invariant under $\{A\}$, for

$$A \sum_n u_n P_{nr} = \sum_{n,m} u_m U_{mn}^i(A) P_{nr}$$
$$= \sum_{s,m} u_m P_{ms} U_{sr}^j(A), \quad \text{from (A16).}$$

But S^i is irreducible under $\{A\}$, therefore this sub-space is either identical to S^i, in which case $n_j \geqslant n_i$, or the sub-space is the zero space, in which case $P \equiv 0$.

In order to prove, on the other hand, that (A16) implies either $n_i \geqslant n_j$ or $P \equiv 0$ it is convenient to define a space S^j, basis $\{v_r\}$, $r = 1 \ldots n_j$, and a set of operators $\{\mathscr{A}\}$, isomorphic to the given set $\{A\}$, which operate from the right on the vectors $\{v_r\}$, such that

$$v_r \mathscr{A} = \sum_s U_{rs}^j(A) v_s. \qquad \ldots\ldots(A17)$$

With this convention we define the sub-space of S^j spanned by the set of n_i vectors

$$\sum_r P_{nr} v_r.$$

This sub-space is invariant under $\{\mathscr{A}\}$, for

$$\sum_r P_{nr} v_r \mathscr{A} = \sum_{r,s} P_{nr} U_{rs}^j(A) v_s$$
$$= \sum_{m,s} U_{nm}^i(A) P_{ms} v_s, \quad \text{from (A16).}$$

But S^j is irreducible under $\{\mathscr{A}\}$, therefore this sub-space is either identical to S^j, in which case $n_i \geqslant n_j$, or the sub-space is the zero space, in which case $P \equiv 0$.

Combining our conclusions, either $P \equiv 0$, or $n_i = n_j$ and determinant $P \neq 0$, since in this latter case the row vectors of P are linearly independent (and so are the column vectors). The conclusion $n_i = n_j$, $\det P \neq 0$, is untenable however, for this would imply that the irreducible representations $\{U^i\}$ and $\{U^j\}$ were equivalent, contrary to hypothesis. It follows that (A16) implies $P \equiv 0$.

APPENDIX II

Theorems on cosets.

Let G denote a finite group of order N, and $H \equiv \{h_r\}$, $r = 1 \ldots m$, a sub-group of G of order m. We propose to show that G can be decomposed into N/m sets, each of m elements, called left-cosets, of the form $\{b_\alpha h_r\}$, where the N/m elements b_α of G are called the left-coset representatives of G with respect to the sub-group H. (This is true also for right-cosets.) We show first, given two left-cosets $\{b_\alpha h_r\}$ and $\{b_\beta h_s\}$, that *either* $\{b_\alpha h_r\}$ and $\{b_\beta h_s\}$ contain exactly the same elements of G, *or* $\{b_\alpha h_r\}$ and $\{b_\beta h_s\}$ contain no elements in common; for suppose $b_\alpha h_s = b_\beta h_t$ is common to $\{b_\alpha h_r\}$ and $\{b_\beta h_s\}$, then $\{b_\alpha h_r\} = \{b_\alpha h_s h_r\} = \{b_\beta h_t h_r\} = \{b_\beta h_s\}$.

Conversely, the cosets $\{b_\alpha h_r\}$ and $\{b_\beta h_s\}$ contain no elements in common if b_β is not contained in $\{b_\alpha h_r\}$, for b_β is an element of $\{b_\beta h_s\}$. It follows, since the group G is finite, that we shall achieve the desired decomposition of G if we choose the ordered set of N/m coset representatives $\{b_\alpha\}$, such that

$$b_\gamma \notin \{b_\beta h_r\} \quad \text{for all } \beta < \gamma.$$

It is convenient to choose $b_1 = I$, the common identity element of H and G. With this convention, the first coset $\{b_1 h_r\}$ is identical to H.

Now let S_H, basis $\{u_p\}$, denote a space irreducible under H with character $\{\chi^i(h_r)\}$. Furthermore, let S denote the space spanned by the vectors

$$\{b_\alpha u_p\}, \quad \alpha = 1 \dots N/m, \quad p = 1 \dots n_i.$$

The space S so defined is invariant under G, for we can decompose G in the form $G \equiv \{b_\beta h_r\}$, and then

$$g(b_\alpha u_p) = (g b_\alpha) u_p = (b_\beta h_r) u_p = \sum_q (b_\beta u_q) U_{qp}{}^i(h_r),$$

where b_β and h_r are defined by $g b_\alpha = b_\beta h_r$.

Corresponding to each coset-representative b_α, this space S contains a sub-space S_α, defined by the basis $\{b_\alpha u_p\}$, where $S_1 \equiv S_H$. The sub-space S_α is irreducible under the sub-group $H_\alpha = \{b_\alpha h_r b_\alpha{}^{-1}\}$, where $H_1 = H$, for

$$(b_\alpha h_r b_\alpha{}^{-1})(b_\alpha u_p) = \sum_q (b_\alpha u_q) U_{qp}{}^i(h_r).$$

Condition that S *be irreducible under* G.

The space S is the sum of the sub-spaces S_α. We are concerned with the case where no two of these sub-spaces possess a non-zero vector in common. We can then take as our basis in S the set of vectors $\{b_\alpha u_p\}$, $p = 1 \dots n_i$, $\alpha = 1 \dots N/m$. Now consider the condition

$$\sum_g \chi(g) \chi^*(g) = N,$$

that S shall be irreducible under G. Using the basis $\{b_\alpha u_r\}$ in S, we can decompose the spur $\chi(g)$ of $g \in G$ into contributions $\chi_\alpha(g)$ from each of the sub-spaces S_α; that is $\chi(g) = \Sigma_\alpha \chi_\alpha(g)$. Then

$$\sum_g \chi(g) \chi^*(g) = \sum_g \Big[\sum_\alpha \chi_\alpha(g)\Big] \Big[\sum_\beta \chi^*{}_\beta(g)\Big] = \sum_{\alpha,\beta} \sum_g \chi_\alpha(g) \chi^*{}_\beta(g)$$
$$= \sum_\alpha \sum_g |\chi_\alpha(g)|^2 + \sum_{\alpha \neq \beta} \sum_g \chi_\alpha(g) \chi^*{}_\beta(g).$$

Now from our relations

$$g(b_\alpha u_p) = \sum_q (b_\beta u_q) U_{qp}{}^i(h_r), \quad \text{where } g b_\alpha = b_\beta h_r,$$

for the matrix representative of g relative to the basis $\{b_\alpha u_p\}$ in S, it is evident that $\chi_\alpha(g) = \chi^i(h_r)$ if g is such that $g = b_\alpha h_r b_\alpha{}^{-1}$; that is, $g \in H_\alpha$, but $\chi_\alpha(g) = 0$ if $g \notin H_\alpha$. It follows that

$$\sum_\alpha \sum_g |\chi_\alpha(g)|^2 = \sum_\alpha \sum_r |\chi^i(h_r)|^2 = \sum_\alpha m = N,$$

since χ^i is an irreducible character of H. Furthermore, the only non-zero terms in the sum $\Sigma_g \chi_\alpha(g) \chi^*{}_\beta(g)$, correspond to elements of G common to H_α and H_β. Now the elements common to two groups form a common sub-group, for suppose ξ and η are elements common to H_α and H_β, then ξ^{-1}, η^{-1}, $\xi\eta$, $\eta\xi$, are common to H_α and H_β, while the unit elements of H_α and H_β are identical, since H_α and H_β are sub-groups of G. Let $H'_{\alpha\beta} \equiv \{\xi\}$ denote the sub-group common to H_α and H_β, then

$$\sum_g \chi_\alpha(g) \chi^*{}_\beta(g) = \sum_\xi \chi_\alpha(\xi) \chi^*{}_\beta(\xi).$$

It follows that the space S is irreducible under G if

$$\sum_{\xi} \chi_{\alpha}(\xi)\chi^{*}_{\beta}(\xi) = 0 \quad \text{for all } \alpha \neq \beta.$$

This curious condition is satisfied in the case of crystal space-groups, Part 2, §5.

Reducibility of S *under* G.

If the space S considered above is reducible under G, we can analyse the reducible character $\{\chi(g)\}$ of S as follows. Let $\{\chi'(g)\}$ denote an irreducible character of G, then the number of times χ' occurs in the decomposition of χ is

$$\begin{aligned} c' &= \frac{1}{N}\sum_{g}\chi(g)\chi'(g^{-1}), \\ &= \frac{1}{N}\sum_{\alpha}\sum_{g}\chi_{\alpha}(g)\chi'(g^{-1}) \\ &= \frac{1}{N}\sum_{\alpha}\sum_{r}\chi^{i}(h_r)\chi'(b_{\alpha}h_r b_{\alpha}^{-1}) \\ &= \frac{1}{m}\sum_{r}\chi^{i}(h_r)\chi'(h_r), \end{aligned}$$

since $\chi'(b_{\alpha}h_r b_{\alpha}^{-1}) = \chi'(h_r)$, h_r and $b_{\alpha}h_r b_{\alpha}^{-1}$ being in the same class of G. This result shows that the number of times $\{\chi'(g)\}$ occurs in $\{\chi(g)\}$ is equal to the number of times $\{\chi^{i}(h_r)\}$ occurs in $\{\chi'(h_r)\}$.

APPENDIX III

The following books and articles are suggested for parallel or further reading :

Modern algebra.

Albert (1937), Halmos (1947), Littlewood (1950), Von Neumann (1950).

The crystal point groups and space groups.

Bethe (1929), Kittel (1953), Koster (1956), Lomont (1959), Margenau and Murphy (1943), Murnaghan (1938).

The permutation groups.

Corson (1951), Ledermann (1949), Littlewood (1950), Murnaghan (1938), Rutherford (1948), Weyl (1931, 1939), Wigner (1959).

The rotation groups.

Corson (1951), Eckart (1930), Edmonds (1957), Fano and Racah (1959), Littlewood (1950), Murnaghan (1938), Weyl (1931, 1939), Wigner (1959).

References

Albert, A. A., 1937, *Modern Higher Algebra* (Chicago : University of Chicago Press).
Bell, J. S., 1955, *Proc. Roy. Soc.* A, **231**, 479.
Bethe, H., 1929, *Ann. Phys., Lpz.*, **3**, 133.
Bouckaert, L. P., Smoluchowski, R., and Wigner, E., 1936, *Phys. Rev.*, **50**, 58.
Corben, H. C., and Stehle, P., 1950, *Classical Mechanics* (New York : John Wiley).

CORSON, E. M., 1951, *Perturbation Methods in the Quantum Mechanics of n-electron Systems* (New York : Hafner; London: Blackie).

DIRAC, P. A. M., 1947, *The Principles of Quantum Mechanics*, 3rd Edn (Oxford : University Press).

ECKART, C., 1930, *Rev. Mod. Phys.*, **2**, 344.

EDMONDS, A. R., 1957, *Angular Momentum in Quantum Mechanics* (Princeton : University Press).

ELLIOTT, J. P., JUDD, B. R., and RUNCIMAN, W. A., 1957, *Proc. Roy. Soc.* A, **240**, 509.

ELLIOTT, R. J., 1954, *Phys. Rev.*, **96**, 280.

FANO, U., and RACAH, G., 1959, *Irreducible Tensorial Sets* (New York : Academic Press).

GABRIEL, J. R., 1960, *Proc. Camb. Phil. Soc.*, in the press.

HALMOS, P. R., 1947, *Finite Dimensional Vector Spaces* (Princeton : University Press).

HERRING, C., 1937, *Phys. Rev.*, **52**, 361.

KITTEL, C., 1953, *Introduction to Solid-state Physics* (New York : John Wiley).

KOSTER, G. F., 1956, *Massachusetts Institute of Technology Technical Report*, No. 8.

KRAMERS, H. A., 1930, *Proc. Acad. Sci. Amst.*, **32**, 959.

LEDERMANN, W., 1949, *Introduction to the Theory of Finite Groups* (Edinburgh : Oliver and Boyd).

LITTLEWOOD, D. E., 1950, *The Theory of Group Characters*, 2nd Edn (Oxford : University Press).

LOMONT, J. S., 1959, *Applications of Finite Groups* (New York : Academic Press).

MCWEENY, R., 1955, *Massachusetts Institute of Technology Technical Report*, No. 7.

MARGENAU, H., and MURPHY, G. M., 1943, *The Mathematics of Physics and Chemistry* (New York : Van Nostrand).

MATTHEISS, L. F., 1959, *Massachusetts Institute of Technology Progress Report*, April 1959.

MURNAGHAN, F. D., 1938, *The Theory of Group Representations* (Baltimore : Johns Hopkins Press).

OPECHOWSKI, W., 1940, *Physica*, **7**, 552.

RUTHERFORD, D. E., 1948, *Substitutional Analysis* (Edinburgh : University Press).

SEITZ, F., 1936, *Ann. Math.*, **37**, 17.

VON NEUMANN, J., 1950, *Functional Operators*, Vol. II : *The Geometry of Orthogonal Spaces* (Princeton : University Press).

WEYL, H., 1931, *The Theory of Groups and Quantum Mechanics* (New York : Dover) ; 1939, *The Classical Groups* (Princeton : University Press).

WIGNER, E., 1932, *Nachr. Ges. Wiss. Göttingen*, 546 ; 1959, *Group Theory* (translated by GRIFFIN, J. J.) (New York : Academic Press).

YAMANOUCHI, T., 1935, *Proc. Phys. Math. Soc. Japan*, **17**, 274 ; 1936 a, *Ibid.*, **18**, 10 ; 1936 b, *Ibid.*, **18**, 623.

PHOTOELECTRIC PHOTOMETRY

By H. J. J. BRADDICK
The Physical Laboratories, University of Manchester

CONTENTS

	PAGE
§ 1. Introduction	154
§ 2. Photocells and photocathodes	155
§ 3. Electron multipliers	158
§ 4. Photomultipliers in practice	160
§ 5. Weak light photometry	161
§ 6. Electron pulse counting	165
§ 7. The photocell compared with other photometric detectors	167
§ 8. Photometry of the farther ultra-violet	168
§ 9. Solid state devices	170
§ 10. Examples of special applications of photoelectric photometry	172
10.1. Precision photometry	172
10.2. Multichannel spectroscopy for industrial analysis	172
10.3. Some differential applications	173
Acknowledgments	174
References	174

Abstract. The development of photoelectric devices has occasioned considerable changes in all types of instrument in which intensities are measured or compared. These devices are directly sensitive to flux and the optical systems used with them must be designed accordingly. At very low intensities the photomultiplier cell allows a light flux of only a few quanta per second to be measured and the theoretical limitations and practical arrangements are discussed. In the ultra-violet spectrum, the photoelectric technique must be revised, but efficient detectors for the whole region are known and some are described. Infra-red detection requires recourse to the 'internal' photoelectric effect in semiconductors, and the properties of the most important detectors of the class are briefly described.

A characteristic advantage of photoelectric detectors is that their output is linearly related to the intensity of the incident light and is in a form suitable for electronic data processing by analogue or digital methods. Some examples in industrial and astronomical photometry which make use of this property are described, and there are some notes on precision photoelectric photometry as used in the maintenance and use of standards of radiation.

§ 1. Introduction

Photoelectric devices have been used in photometry for about fifty years (e.g. Elster and Geitel 1910); comparatively recently, and especially since photomultiplier cells have been available, their characteristics have led to radical innovations in the study of radiation. Optical instruments designed for visual or photographic observation are not necessarily best when information is to be extracted from incoming radiation by photoelectric devices, and this fact is now influencing, in particular, the design and choice of spectrometers. When information about light intensity is available in the form of an electrical signal it becomes possible to transform it in various useful ways; as a simple example, two intensities may be measured alternately with any appropriate frequency and their equality or inequality observed with great accuracy. Variations common to both intensities or slow variations in the characteristics of the detector can be eliminated.

In all the devices we shall consider the primary process is the absorption of quanta with liberation of electrons. If n quanta are incident per second and if, on the average, one quantum produces ϵ 'liberation events' then, over a measurement period t, the standard deviation of the number of events is a fraction $(1/n\epsilon t)^{1/2}$ of the number. This represents the natural limit to the accuracy of photoelectric measurement of a steady intensity. The signal/noise ratio is $(n\epsilon t)^{1/2}$, and the 'equivalent quantum efficiency' of a process could be defined as the value of ϵ which gives the observed signal/noise ratio, even in those cases where no other knowledge of the primary efficiency is available (Rose 1948, Fellgett 1955).†

It is clear that the quantum efficiency is fundamental to the precision of measurement; ϵ must always be less than unity and the signal/noise ratio is not affected by subsequent (perfect) amplification. In general the ratio is degraded by further fluctuations introduced in the measuring system, but in certain cases of weak light photometry the quantum limit is closely approached.

§2. Photocells and Photocathodes

In the simplest emission photocells electrons from the cathode are directly collected (vacuum photocell). Internal amplification may be obtained by filling the cell with chemically inert gas at a suitable pressure and applying enough accelerating field to the electrons to produce ionization by collision. An amplification of the order of one hundred may be obtained in this way, but the amplification depends critically on the applied voltage and on minor changes of gas pressure. Gas-filled cells were once used in astronomical photometry but they are not now important for quantitative work and find their chief application in certain industrial applications where an 'all or none' response is all that is required.

In the so-called photomultiplier cells the electrons, after acceleration, fall on a prepared surface (dynode) which gives a high yield of secondary electrons. An amplification of about 3 to 5 is achieved in a single stage, and multiplier cells with 9 to 19 stages are in common use.

The efficiency of the cathode is fundamental. In pure metals the extraction of electrons, at least for incident quantum energies near the threshold, takes place only in a very thin layer where the surface fields are effective. There are technical difficulties both in experimental measurement and in quantum-mechanical calculation, but there is a reasonable agreement that for metals in general and alkali metals in particular ϵ is of the order of 10^{-4} and reaches a maximum for a frequency of about 1·5 times the threshold frequency (for references see Simon and Suhrmann 1958).

Pure metals are not very important in general photometric technique, though potassium vacuum cells were formerly used in precision photometry (§ 10), and cells for measuring ultra-violet radiation for meteorological and biological purposes have been constructed with cathodes of sodium or zirconium (Coblentz 1946, other references in Simon and Suhrmann 1958), since they are not sensitive to stray radiation of relatively long wavelength. A number of metals have been used as cathodes in the photometry of the far ultra-violet (§ 8). The first important method of enhancing the sensitivity of an alkali metal cathode consisted in a controlled treatment with a hydrogen glow discharge. The maximum of ϵ rises by a factor

† The efficiency of the eye and the photographic plate will be compared with that of photoelectric devices in § 7.

10–100, and the threshold of sensitivity is displaced to longer wavelengths. The sensitivity of these hydride cathodes is ascribed to the simultaneous presence of alkali metal and alkali hydride, and in fact very many compounds on which alkali metals are adsorbed give photocathodes of relatively high sensitivity. A photocathode of technical importance is obtained by exposing an oxidized surface of silver to treatment with caesium vapour. The sensitivity of this cathode (Type S1) corresponds to a maximum ϵ value of about 1% between 7000 Å and 8000 Å and its sensitivity extends in the red to nearly 12000 Å (1·2 μ).

Many of the most important modern cathodes are developed from the discovery by Görlich (1936) that a highly sensitive surface is obtained by the action of alkali (especially caesium) vapour on a layer of antimony or bismuth. The activated layer contains compounds of the type Cs_3Sb: these are probably intrinsic semiconductors in which electrons can be extracted by light from the lattice and not merely from the surface layers. The cathode films are semi-transparent (absorbing for example 30% of the incident light) and the cathode is often deposited on a transparent wall or window of the tube so that the electrons are liberated on the side remote from the incident light. The efficiency of these semi-transparent cathodes is not very different from that of caesium–antimony cathodes deposited on a solid metal substrate illuminated from the front, presumably because the depth from which electrons can escape is limited. The efficiency can be somewhat improved by depositing the semi-transparent cathode on a highly reflecting surface and arranging to extract the electrons from the side on which the light is incident.

The caesium–antimony and caesium–bismuth cathodes have been extensively developed by the manufacturers, but details of these developments have not been fully published; they involve the controlled addition of oxygen and in some cases alkali metals and metallic silver. The processes are in practice not fully controlled, so that when a tube is required to have the best possible performance in some particular—e.g. absolute current sensitivity, low dark current—selection among a large stock is very advantageous.

The published sensitivities ascribed to a given type of cathode vary considerably. Engstrom, Staudenheimer and Glover of RCA (1952) have published statistical data on a batch of production multiplier tubes intended for nuclear scintillation work in which the light emitted by the phosphor was to be blue. They show that the sensitivity to white light varied within the batch by a factor of more than two, and the spectral characteristics also varied so much that there was only a rough correlation between the sensitivity to white light and the sensitivity to blue light.

It is known that caesium–antimony surfaces are obtained which have peak quantum efficiencies near 30%. A recent development is the multi-alkali cathode with the general composition M_3Sb, when M is now a mixture of two or more alkali metals (Sommer 1955). According to Sommer the most sensitive cathodes contain sodium and potassium in the atomic ratio 2 : 1 with only a little caesium. The tri-alkali cathode has a higher peak quantum efficiency near 4200 Å than most caesium–antimony cathodes (30% is regularly obtained), and its efficiency is much better maintained in the long-wave part of the spectrum. The cathode has a useful sensitivity at 8500 Å and its sensitivity to standard tungsten light (black-body radiation at 2870°K) is nearly three times as great as that of a caesium–antimony surface.

Figure 1 shows the quantum efficiency of a number of commercial (E.M.I.) cathodes ; similar figures are obtained from data published by other manufacturers. Unfortunately, in commercial publications, the sensitivity of cathodes is often given in microamperes per lumen (usually to tungsten light : effectively black-body radiation at a quoted temperature). This quantity corresponds, of course,

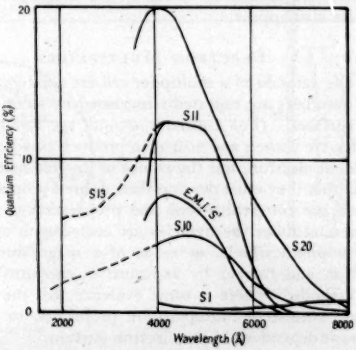


Figure 1. Published quantum efficiency of cathodes (E.M.I.). S1 : Cs–O–Ag ; S10 : Cs–O–Ag–Bi ; S11 : Cs–Sb–O ; S13 : S11 in quartz. E.M.I. 'S' Cs–Sb with suppressed red response.

Table 1. Commercial Cathode Designations

S1	Cs–O–Ag	Denominated S by Cintel, C by Mullard, CM by G.E.C.
S3	Cs–O–Ag–Bi	Denominated B by Cintel. Gives response over whole visual spectrum.
S4	Cs–Sb (metal backing)	Denominated A by Cintel, Mullard, CA by G.E.C. Response roughly similar to S11, figure 1.
S5	Cs–Sb (metal backing and u.v. glass bulb)	Response same as S4, but extends to about 2500 Å.
S8	(semi-transparent)	Very similar to S3.
S9	Cs–Sb	Similar to S4, but modified by absorption of light in film.
S10	Cs–O–Ag–Bi (semi-transparent)	
S11	Cs–Sb–O (semi-transparent)	Efficiency higher than S9, spectral characteristics very similar to S4.
S13	Cs–Sb–O (semi-transparent on quartz window)	Response same as S11, but extends below 2000 Å.
S17		RCA high efficiency cathode on reflecting backing. Spectral response 2900–6200 Å, with maximum about 5000 Å.
S19		Quartz bulb for ultra-violet.
S20	Tri-alkali	

The E.M.I. 'S' cathode is Cs–Sb treated to give low dark current at room temperature at the expense of some sensitivity.

to an easily measured property of the tube, but it is only useful for comparing cathodes of closely similar spectral characteristics, and it can be grossly misleading in comparing, for example, a red-sensitive Ag–O–Cs cell with an Sb–Cs cell having its maximum efficiency in the violet. Table 1 gives the code symbols which are regularly used by American manufacturers to describe the cathode characteristics of their tubes; these are sometimes used in this country, but other less well standardized symbols also appear.

§ 3. Electron Multipliers

Electrons from the cathode of a multiplier cell are accelerated into the dynode system where their numbers are repeatedly increased by secondary emission from successive dynode surfaces. (For a general account see Rodda 1953 and Sharpe 1957.) The dynodes are chosen and coated to produce a useful number of secondaries for each incident electron, but the choice of dynode materials is in practice restricted by the fact that they must develop their desired properties under processing conditions which are compatible with the preparation of an efficient photocathode. In commercial tubes the dynodes are coated with a caesium–antimony surface rather like a photocathode, or made of a magnesium–silver alloy, and activated by oxidation and further by exposure to caesium vapour during the preparation of the cathode. There is some evidence that the caesium–antimony surface has the more consistent multiplication properties, in particular that the amplification is more independent of the electron current.

The geometry of the dynode systems in common use is indicated in figure 2; the electric fields must be such as to accelerate the secondary electrons away from the surface of each dynode. In the systems 2(*a*), 2(*b*) and 2(*c*) the electrons are focused electrostatically, so that the descendants of a particular primary electron form a compact bundle at every stage, in the other systems focusing does not take place. In order to reduce the dark current due to thermionic emission of electrons it would be valuable if the focusing arrangements could be made to collect electrons into the multiplier only from a small cathode area and not from considerable areas of caesiated materials at or near cathode potential. This does not yet seem to be achieved in any commercial tube though it is claimed for tubes made by Lallemand on a laboratory scale. In most photometric experiments the different characteristics of focusing and non-focusing dynode systems are not very important; they are significant in nuclear scintillation counters where the spread in the time of collection of the electrons arising from a particular event decides the shape of the output pulse and the possibilities of time discrimination. Focused multipliers are considerably better in this respect than unfocused.

The effective gain of a multiplier stage depends on the secondary emission ratio and the efficiency with which the electrons are collected. The average stage gain usually lies in the region 3 to 5 at the normal working voltage, and at least with Sb–Cs dynodes, the variation of stage gain with voltage is given by

$$\frac{d(\log \text{ current gain})}{d(\log \text{ voltage})} = 0{\cdot}7$$

in the neighbourhood of the working voltage. If the voltage applied to a multiplier of several stages is increased, instability eventually appears, and this may be

ascribed to the extraction of electrons from the sensitive surfaces at the cathode end of the system by radiation or positive ions leaking back from the region of high electron current near the collector. Precautions are taken in the design of

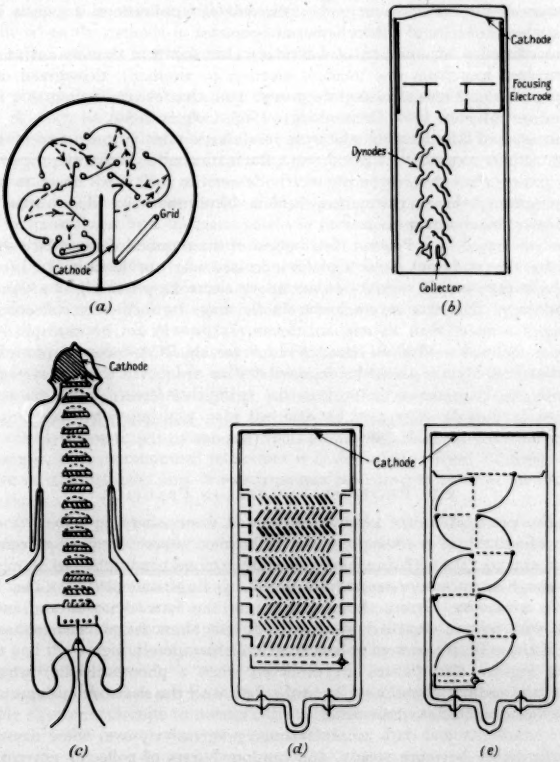


Figure 2. Schematic sketches of dynode systems, (*a*) compact focused system (Rajchman, RCA), (*b*) linear focused system (RCA), (*c*) Venetian blind with focusing grids (Lallemand 1949), (*d*) Venetian blind with flat grids (E.M.I.), (*e*) box and grid system (DuMont–E.M.I.).

tubes to avoid this feedback, but it will always appear if the multiplier gain is sufficiently increased, and the effect is likely to vary from tube to tube so that the highest stable amplification obtainable from a given tube design is a rather uncertain quantity.

The linearity of a multiplier system (constancy of gain with varying input current) is as good at relatively low currents as the measurements disclose. Space charge effects appear in the later multiplication stages at higher currents; these effects can be delayed by increasing the voltages applied to the later stages, but for other reasons the collector current in photometric applications is usually kept so low that non-linearity of this type is unimportant.

The efficiency of emission of secondary electrons at a dynode surface varies in a random way from one incident electron to another; the spread deduced from experimental results is slightly greater than that corresponding to a Poisson distribution (Morton 1949, Breitenberger 1955, Colgate 1959).

This spread affects photometry in two ways—the output current from a photomultiplier exposed to light shows a fluctuation noise which is proportionally rather greater than that in the photocathode current and, if we resort to electron pulse counting, the output pulses have a continuous size distribution which complicates the counting.

Analysis based on a Poisson distribution of the number of secondary electrons shows that the ratio r.m.s. noise/signal is increased nearly in the ratio $[1+1/(\sigma-1)]^{1/2}$ where σ is the average number of secondary electrons produced at a dynode by each primary. If electrons are lost from the stage by inefficient collection, or if the spread is more than Poissonian, the term $1/(\sigma-1)$ can be multiplied by an empirical factor k. Morton found $k=1{\cdot}5$ for an RCA focused multiplier; a somewhat larger factor would be expected for an unfocused dynode system, and, of course, the degradation is increased by failures of focusing due, for example, to incorrect dynode voltages. In practical photomultipliers correctly used the noise/signal ratio is about 25% more than that due to the photocurrent.

§ 4. Photomultipliers in Practice

In the practical use of photomultipliers it is necessary to attend to certain effects whose origin is obscure. The cells often show 'fatigue' effects; the collector current falls with time under constant external conditions and illumination. Some rough quantitative results are given by Engstrom (1947). The fatigue probably arises by electron bombardment of the later dynodes and increases rapidly with output current; most cells do not show fatigue changes as great as 1% if the collector current is less than 1 microampere.†

Less regular disturbances are observed when a photomultiplier which has been kept in ordinary light is set up in the dark with the electrode voltages applied and a measuring circuit connected.‡

The sensitivity and dark current change progressively over some days before the performance becomes steady, and random bursts of collector current occur during this period. Some of these effects are probably ascribable to the redistribution of charges on the envelope and other insulating surfaces within the tube;

† Lenouvel and Daguillon (1954) observed 'fatigue' in a Lallemand 19-stage multiplier at collector currents above 4×10^{-8} amp; they ascribed these to low conductivity at the dynode surfaces. Their results probably refer specifically to silver–magnesium dynodes and to Lallemand's special disposition of the collecting electrode in front of the last dynode.

‡ No photocell should be exposed to really strong light, even without applied voltage, nor should the cathode be exposed to temperatures much above room temperatures.

some workers recommend coating the tube with a conducting layer (e.g. graphite) maintained at or near cathode potential.

Large disturbances occur if the tube is touched or comes into contact with objects not at cathode potential, and if it is necessary for optical or other reasons to have supports in contact with the glass, they must be at cathode potential, especially when the cathode is deposited on a wall of the tube. It may be advantageous to use a circuit in which the cathode is ' earthy ' and the collector at a high positive potential.

The potentials of the intermediate dynodes are usually derived from a resistor chain; it may be possible to improve the performance of electron-focusing multipliers by adjusting the potentials of particular dynodes. The behaviour of an electron multiplier is affected by magnetic fields, and focused multipliers are more sensitive to them than unfocused. It is certainly necessary to avoid strong magnetic fields, and even the earth's field may disturb a focused multiplier and should be neutralized by a mumetal shield in critical cases, particularly when the tube must be moved during measurements.

The effective sensitivity of many photomultipliers is found to vary greatly if a small spot of light is moved across the cathode surface. This variation is largely due to the different efficiency with which electrons are collected from different parts of the cathode; it is especially marked in the RCA 1P21 and similar cells (data given by Saunderson, Caldecourt and Peterson (1945) show that the sensitivity changes by a factor of 20 in a few millimetres); it is present, though not so strongly, in ' Venetian blind ' photomultipliers. It is therefore necessary in photometric applications to distribute light over the cathode in a way which does not vary uncontrollably from time to time and appropriate optical systems must be used. For example, with astronomical telescopes it is desirable to form an image of the objective on the cathode, and in spectroscopic instruments to form an image of the slit there.

From the data given above on the variation of gain with voltage, it is clear that a well stabilized electric supply is needed if constant gain is to be obtained from the multiplier (e.g. 0·1% variation in voltage produces nearly 0·8% variation in the gain of an eleven-stage multiplier). Some workers use batteries of small cells; in most circumstances a mains-driven power unit is a less troublesome solution, and the type is to be recommended in which a high-frequency oscillator feeds a rectifier and the amplitude of oscillation is controlled by a feedback loop containing a voltage reference tube.

Photomultiplier tubes can be used for general photometry, with galvanometers or simple electronic circuits to measure the collector current, but they have special importance in the measurement of quite weak light.

§ 5. Weak Light Photometry

Many of the developments in the photoelectric photometry of weak light have been made for astronomical purposes. This is not surprising since a sixth magnitude star, just visible to the naked eye, yields about 4000 visible quanta per cm^2 per sec at sea level, and increasing the collecting area is a costly business. The 200-inch telescope collects about 300 quanta per second from a 22nd magnitude star, which is near its photographic limit, and for most astronomical purposes, e.g.

spectroscopy, colorimetry, polarimetry, only a fraction of the incident flux can be used in one of the information channels. There are occasions in laboratory spectrometry (e.g. micro-spectrophotometry, the analysis of rare isotopes) where the flux of light is comparable with that available in astronomical work.

In the visible, near infra-red, and ultra-violet the multiplier cell is unrivalled; it will be shown below that at very low intensities there are severe difficulties in using a vacuum cell with external amplification. A very important limit to sensitivity comes from the 'dark noise' of the photocathode. All multiplier tubes give an appreciable collector current in the dark and the fluctuations of the dark current are superposed on the signal. Since the origin of the dark current and dark noise is not fully known, these fluctuations cannot be calculated with confidence, and experiment is required.

The dark current is primarily due to thermionic emission; dark current and dark noise can be reduced by cooling the cell. The pioneer work reported by Engstrom (1947) and RCA catalogue data on recent tubes both indicate that the noise output (square of fluctuation voltage) of a caesium–antimony cathode decreases by a factor of 100 for a temperature reduction of 80°c about room temperature, and the cathode current presumably decreases by a similar factor. This reduction corresponds to a work function (about 0·3 volt) far below that deduced from the photoelectric threshold (about 2 volts), so that a simple explanation does not suffice. It has been suggested that the photocathode surface is heterogeneous and does not follow Richardson's law. The temperature effect probably varies appreciably from tube to tube. Electron counting experiments at Manchester on E.M.I. tubes have usually shown a considerably larger reduction in the current than that given above. Changes in the photoelectric efficiency of cathodes with cooling are small and rather uncertain, though it is found that the conductivity of caesium–antimony layers decreases at low temperatures, and it is recommended that a transparent conducting layer of tin oxide be interposed between glass and cathode layers in cells which are to be used at liquid air temperatures (Harper and Choyke 1956). The conductivity of the tri-alkali cathode is said to be better maintained at low temperatures.

Refrigeration is, then, important in measuring weak light, though there are technical difficulties in avoiding the condensation of moisture on cells and windows, and tubes are sometimes damaged by thermal shock. Some workers use liquid air for cooling, but carbon dioxide snow is a very convenient cooling agent and with some cells at least it reduces the dark emission to a very few electrons per cm^2 per second. Red-sensitive (Cs–O–Ag) cathodes have a thermionic emission at ordinary temperatures which is a few hundred times greater than that of Cs–Sb, and cooling is even more important when weak light is measured in the red part of the spectrum. Red-sensitive photomultipliers have only recently become generally available; their use in astronomy has been discussed by Kron (1958) who compares a number of cells, all with Cs–O–Ag cathodes and Ag–Mg dynodes. The cathode sensitivities of these cells varied by a factor of about 4; the thermionic emission inferred from the dark noise was about 10^{-18} amp (a few electrons per second) at dry ice temperature and about 10^{-11} amp at room temperature.

In the photometry of weak light, the superiority of an electron multiplier lies in the fact that it introduces much less noise than other forms of amplification. If we

consider the system of figure 3, where a cell with internal current magnification M is coupled to an amplifying circuit giving an output $g(\nu)$ for unit input at frequency ν, the noise output due to fluctuations in the photocurrent i_p may be written

$$\bar{v}^2 = 2ei_p M^2 R^2 K \quad \text{where } K \text{ denotes } \int_0^\infty \frac{g^2(\nu)}{1+4\pi^2\nu^2 R^2 C^2}\, d\nu$$

and the noise output due to grid current in the first amplifying stage is

$$\bar{v}_g^2 = 2ei_g R^2 K$$

while the coupling resistor contributes $4kTRK$. It is clear that with a non-multiplying phototube, the effective sensitivity of the system is limited by the grid current of the best available electrometer valve (about 10^{-15} amp or about 6000

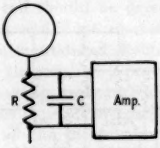


Figure 3.

electrons/sec). At this current the resistor thermal noise will be dominant unless the resistor R is large, and the time constant of the input circuit then becomes inconveniently long. It has been suggested that the electrometer tube should be followed by an amplifier with a rising frequency characteristic (H. L. Johnson 1948) but the whole situation is changed if a multiplying cell is available, and a comparatively simple electronic arrangement following the cell is adequate down to the lowest intensities.

Lallemand (1953) has used a specially made 19-stage multiplier with a very efficient cathode (ice cooled) and measured the collector current with a sensitive galvanometer; he reaches in this arrangement the fluctuation limit of the cathode. Most other users follow the multiplier with an electronic measuring circuit, and three main types may be distinguished:

(*a*) d.c. amplifying circuits

(*b*) a.c. amplifiers using interrupted light

(*c*) electron pulse counting.

In each case the fluctuations of the photocurrent must be averaged out by including a circuit with a long integrating time constant (narrow frequency bandwidth) and it is desirable that this time constant be adjustable to suit the light intensity to be measured.

In the d.c. method it is apparent from the analysis given above that an electrometer valve should be used to measure the collector current, though a specially low grid current is not usually required to reach the theoretical limit of the photomultiplier cell. A circuit of the type shown in figure 4 is very generally applicable.

If the light incident on the photocathode is interrupted by a periodic chopper, the alternating component of the electrical output can be amplified by a multistage

a.c. amplifier. The bandwidth of this amplifier can be reduced by tuned circuits (practically to about 1 c/s). The amplifier can be followed by a rectifier, preferably a synchronous rectifier operating in phase with the light chopper, and time integration can then be effected by d.c. condenser circuits or by electro-mechanical integrators such as the Velodyne. This a.c. amplification system has considerable practical merit ; there is now no question of zero drift in the amplifier, the amplifier gain can be accurately stabilized by heavy negative feedback, and many kinds of parasitic electrical disturbance are effectively neutralized by the synchronous rectifier. Similarly, stray light on the photocell, not chopped at the correct frequency, has no direct effect on the output signal, though it does add to the fluctuation noise.

Since the photocell collector can be connected to the rest of the equipment by a condenser there is no difficulty in having the collector at high potential, so that

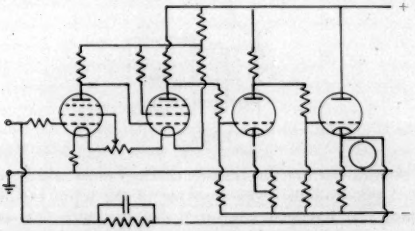


Figure 4. A typical d.c. amplifier for use with photocell or photomultiplier. Note : (*a*) filaments of electrometer and following valve supplied from stabilized h.t. line, (*b*) feedback from output to input over high resistance with capacitor to control time constant.

the photocathode can be at earth potential. The valve amplifiers used hitherto could be replaced by transistor amplifiers without much difficulty, a screened transformer of high ratio being used to couple the photocell to the comparatively low input impedance of a transistor stage.

The system of measurement with chopped light has, however, the property that light only reaches the photocathode for half the time of observation, and the rate of collecting information is correspondingly lowered. At very low light levels, particularly in astronomy, this is a decisive disadvantage unless the alternate intervals can be used for some good purpose, as, for example, if the photometer can be made to look alternately at a star and at the sky background, which in the case of a faint object has an intensity comparable with the star and must be observed with comparable accuracy.

A similar situation arises when the dark current of a cell is comparable with the photoelectric current ; time must then be spent in measuring the dark current and the a.c. method does this automatically and without loss. There are certain special problems in which two or more light sources must be measured simultaneously ; this can be done with a single photocell if orthogonal modulation patterns (e.g. chopping at different frequencies) can be impressed on the different lights and the

output signals unscrambled by appropriate synchronous rectifiers. In absorption spectroscopy, polarimetry, and some other studies, the information required involves essentially the comparison of two intensities ; if these can be made to fall alternately on the cell, the a.c. method can be used with great advantage, particularly if the intensities can be adjusted to equality by means of a calibrated optical attenuator so that the method becomes a ' null ' one.

A.C. amplification with chopped light is a good method to use with non-multiplying photocells, where the light intensity is never so low in practical cases that the time available for observation is a limiting factor. The high and stable amplification required with such cells can well be provided by the a.c. method.

§6. Electron Pulse Counting

A simple calculation shows that the emission of single electrons from the photocathode of a multiplier cell should be detectable ; a charge multiplication factor of 10^6 is easily attainable and if we assume 16 pF as the capacitance of the collector system, the voltage pulse produced should have an amplitude of 10^{-2} volt, which is large compared with the noise of a practical amplifier input circuit (e.g. about 10^{-4} volt r.m.s. for 1 megohm resistance and 1 megacycle bandwidth).

If we neglect temporarily any parasitic or ' dark ' current it appears that a simple counting system can collect all the information which appears in the cathode emission and hence all the information which can be obtained from the incident light by a particular cathode.

Summation of output over an arbitrary time of observation is obtained automatically and the information appears in a digital form which is convenient for certain methods of subsequent processing. The method is best adapted for weak light, as in astronomical applications, and it is not very easy to deal with counting rates faster than a few thousand random events per second. It is likely that the method will find increasing use in non-astronomical photometry, and optical attenuation (e.g. by filters of known transmission) could then be used to assist comparison of stronger and weaker intensities.

Any counting circuit must deal only with pulses greater than a certain critical size ; this is usually set intentionally by means of a discriminator circuit, but discrimination is in all cases inevitable since circuit noise must not produce an important number of counts, and the sensitivity of the system must be limited accordingly.

Because of the nature of the multiplication process, single electrons from the photocathode give output pulses of varying sizes. If therefore the discriminator bias is varied (under constant illumination) the pulse rate varies continuously. Accurate photometry therefore depends on a constant level of discrimination and constant amplification within the cell. This requires a well regulated power supply, such as that described in §4, and a well designed discriminator circuit ; in particular the variable internal potentials which appear in valves must be removed as far as possible by using balanced circuits for the discriminator and stabilizing the heater supplies, as well as the reference voltage.

Pulse height discrimination leads, however, to an advantage characteristic of pulse counting photometry. It is found possible to choose a discriminator bias setting for which the ratio of signal to dark background is higher than the ratio of

the signal current at collector to the background current at collector.† This is due partly to differences between the pulse height distributions of the ' dark ' pulses and those due to illumination of the cathode, and partly to the fact that ohmic leakage currents make no contribution to the pulse rate.

It is found that the output of photomultipliers contains occasional enormous pulses of uncertain origin; these contribute proportionately less to a pulse count than to a current measurement, but they require attention in the design of the pulse amplifier to avoid the fault of ' blocking ' for an appreciable time after an

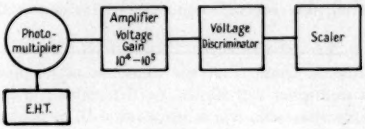


Figure 5

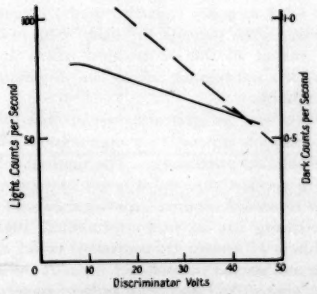


Figure 6. ' Dark ' and ' illuminated ' counts from an electron pulse counting system (photomultiplier cooled with solid CO_2).

overloading pulse. Several astronomical pulse counting photometers have been described (e.g. Yates 1948, Redman and Yates 1953, Blitzstein 1953, 1958) and the principle has been applied to a microspectrophotometer (Thornburg 1956). It might well be advantageous in other problems in physics, chemistry, and biology involving measurement of weak light. Figure 5 gives a block diagram of the apparatus involved and figure 6, taken from a Manchester thesis by N. Woolf, gives an indication of the characteristics of the pulses obtained. These results were obtained with an E.M.I. photomultiplier type 6094A with a 1 cm diameter caesium–antimony cathode, cooled with solid carbon dioxide. The measured efficiency of the photocathode for blue light was about 10% and it appears that a flux of only a few quanta per second could be measured.

† This result has been established by Engstrom (1947) and Blitzstein (1953), and confirmed in much unpublished work in Manchester.

§ 7. The Photocell compared with other Photometric Detectors

It seems useful to compare the photocell with the human eye and the photographic plate, which are its rivals in practical photometry. The eye can only be used as a brightness comparator ; it can in favourable conditions attain a consistency to the order 1%, but the accuracy of comparison at low intensities is much worse than this. The photographic plate has the important advantages of simplicity and long-term storage of information, but it is only an intermediary, since the blackening of the plate must be evaluated by some other photometric method. We shall make quantitative comparisons which apply specially to the photometry of weak light.

In § 1 the equivalent quantum efficiency was introduced as giving a fundamental limitation to the signal/noise ratio of a detector. With the best photocathodes, at the most favourable wavelength, values of $\epsilon = 30\%$ are regularly obtained and efficiencies of up to about 50% have been reported. These values are not attained by any other detector and because the photocell is linear in its response they are available at any required intensity.

Estimates of ϵ for the human eye have been made recently by Fellgett (1958 a) and by Clark Jones (1959). Clark Jones' estimate of 1% was obtained by a sound method but refers to observation at a fairly high background brightness ; a figure of 5% has been obtained, under more doubtful assumptions, for threshold vision.

The equivalent quantum efficiency of photographic emulsions has been estimated by Fellgett (1958 b), using on the one hand the experimental relation between photographic density and exposure, and on the other hand the signal/noise ratio derived from the granularity of the emulsion. For the best emulsions examined ϵ was nearly 1% but this was derived from measurements made at optimum exposure. The photographic plate is far from linear in its response, and the efficiency for long exposures to weak light would be very much less (' reciprocity failure ') unless it were possible to reduce this failure by special treatment of the emulsion or by cooling it during exposure.

As a collector of information the photographic plate has the advantage that many independent channels (10^4–10^6 cm^{-2}) may operate simultaneously on a single plate.†

A photoelectric detector must in general deal with a number of information channels (e.g. points in the profile of a spectral line) by scanning them in a time sequence, though it may, of course, be possible to deploy a number of photocells to deal with a multichannel problem.

In cases, therefore, where the essential problem is to collect information in a limited time, the choice between photocell and photographic plate may depend on the number of picture points (independent information channels) required. From § 1 the time necessary to establish a given signal/noise ratio at a given flux is inversely proportional to ϵ, so that the merit of a given recording system in which n channels are operating simultaneously is given by $n\epsilon$. In the case of the photographic plate, with reciprocity failure, the value of ϵ falls off with the illumination, and in a comparison with the photocell the latter is favoured at low intensities.

Since a photoelectric detector measures simply the photon flux falling on its cathode, any optical system associated with it should be designed to collect as

† Photoelectric image converters in which it is hoped to combine the quantum efficiency of the photocathode with the multichannel property of the photographic plate will not be considered here.

much light as possible from the source, analyse it as required, and transfer it to the photocathode ; the size and shape of the illuminated patch is unimportant and indeed it is made fairly large for technical reasons. The response of a photographic plate to an extended image is more complicated ; if the response were linear the information stored over any area of the plate could be recovered by integrating the photographic density over the area, but in practice quanta spread thinly over a large area are lost to the record. If the intensity is low, and we consider areas large compared with the plate grain, the photometric threshold of the plate is governed by the flux *per unit area*. This means, for example, that a telescope for the photographic photometry of faint stars must be a well-focused instrument and its focal length must be so short that the ' false disc ' produced by diffraction and atmospheric ' seeing ' conditions is not large compared with the inevitable disc produced by the graininess of the plate. The only requirement for photoelectric photometry concerns the separation of the star from its background ; the image must be separated from other discrete sources big enough to disturb the measurement, and the sky light which passes through the limiting diaphragm will degrade the measurement with noise. It may be possible to attack some astronomical problems by using very large collecting systems of fairly low optical quality.

In connection with the spectroscopy of faint sources (those having a low superficial brightness) Jacquinot (1954) showed that the flux through a spectroscopic instrument having a dispersing element of given size and type is inversely proportional to the resolution demanded from the instrument, and he compared the luminous efficiency of different spectroscopes on this basis. It appears that gratings are much superior to prisms, essentially because the higher dispersion attainable allows bigger slits to be used for a given loss in frequency discrimination, and that axially symmetrical interferometers are much better than gratings. But this superiority in *flux* does not involve an increased value of *flux per unit detector area*, so that it is much easier to take advantage of the more luminous dispersing systems using photoelectric and not photographic detectors. It is further necessary for most purposes to use the symmetrical interferometers with time-sequential scanning, so that photoelectric detectors are naturally associated with these new spectroscopic instruments.

§ 8. Photometry of the Farther Ultra-violet

Commercial photocells and photomultipliers are available with quartz envelopes or windows ; the photocathode materials, particularly Cs–Sb, have a high photoelectric efficiency in the ultra-violet, but the transmission of the quartz falls off rather sharply below 2000 Å, becoming negligible at 1550 Å (Dunkelman and Loch 1951). Experimental sealed tubes have been made with windows of synthetic sapphire which transmits well to about 1420 Å. At wavelengths shorter than this, photoelectric detectors are used in vacuum spectrometers, and there has been special interest in photometric detectors selective for particular spectral regions. In particular the Lyman α hydrogen line at 1216 Å is important both in the laboratory and in rocket-borne investigations of sunlight at the top of the atmosphere, and detectors specially sensitive to it have been made. The photoelectric devices used for wavelengths less than 1400 Å have been of three main types : fluorescent

substances used as wavelength converters, photocathodes and photomultipliers which can be used in the vacuum of a spectrometer, and photocounters dependent on gas ionization. Fluorescent converters were investigated by Johnson, Watanabe and Tousey (1951) who found that a film of sodium salicylate crystals, deposited from a methyl alcohol solution, excited by high-frequency radiation, gave out light which was efficiently detected by a photomultiplier having the usual Cs–Sb cathode. The overall efficiency was high and constant over the wavelength range from 2300 Å to 900 Å (Watanabe and Inn 1953).†

Turning to the direct photoelectric method, we find that many metals give photoelectrons with fairly high quantum efficiency under short-wave ultra-violet excitation; this effect was indeed used in many of the pioneer investigations of this part of the spectrum. We are here probably concerned with a 'volume' photoeffect on electrons bound by lattice forces rather than by surface potential gradients; the efficiency and other characteristics of the effect are radically different from the photoelectric effects observed at longer wavelengths. Hinteregger and Watanabe (1953) found that the efficiency of the photoemission from Ni, Pt, W, rises steeply at wavelengths shorter than 1500 Å, to give ϵ values of about 10%. These cathodes were quite insensitive to radiation of longer wavelength; they could be exposed to air without measurable change in their short-wave photoelectric properties. Hinteregger and Watanabe further found that it was possible to distinguish the quantum energy of the exciting radiation by the energy distribution of the emitted electrons, and hence to change the spectral range of the photocell by applying a retarding potential.

It would clearly be useful to construct photomultipliers sensitive to these short wavelengths and this was done by Schwetzoff (1954), who worked especially with Ag–O–Cs surfaces specially treated to destroy their long-wave sensitivity. Schwetzoff also used oxidized Be–Cu surfaces; the secondary emission properties of this substance can be restored by simple heat treatment after exposure to air so that it is practical to connect a photomultiplier to a vacuum spectrograph without an intervening window.

Garton, Webb and Wildy (1957) used a photomultiplier with a tungsten cathode and beryllium copper dynodes to measure optical absorption at 1216 Å. They obtained a current gain of 10^4 with 14 stages and a total voltage of 5 kv; the stage gain was comparable with that observed by Schwetzoff. A similar multiplier was used at much shorter wavelengths (Gyorgy and Kingston 1951) and found many times more sensitive than a photographic plate.

It is reported that windowless photomultipliers are used in Russian rocket instrumentation; several American workers have published descriptions of photoelectric Geiger counters designed for rocket experiments. Chubb and Friedman (1955) made small Geiger counters of conventional form with lithium fluoride windows; with a filling of helium and ethyl formate as a quenching agent, the counters showed a rather low photoelectric sensitivity below 3000 Å, the counts being ascribed to the ordinary photoemission from the chrome–iron cathode. At 1500 Å the quantum efficiency rose steeply with the onset of the inner photoelectric

† An ingenious use of a fluorescent converter has been described by James (1959), who uses a small strip coated with sodium salicylate to convert part of the beam emerging from a monochromator and thus to provide a reference signal in absorption spectrophotometry.

emission; at a still shorter wavelength there was a faster increase in efficiency due to the photo-ionization of the ethyl formate.

It was found that more selective counters could be made if the filling contained certain highly electronegative gases such as nitric oxide or the halogens. The counts due to photoelectrons from the metal wall are then almost entirely suppressed by electron capture and only the sensitivity due to photo-ionization remains.

By varying the gas filling and the filtering effect of the window the spectral region of sensitivity can be in some measure adjusted; with NO it forms a jagged band centred on 1250 Å. Some photosensitivity was observed with chlorine and bromine in the filling, though it was certainly not due to the simple photo-ionization of these elements. The same principle was applied by Brackmann, Fite and Hagen (1958) who made counters filled with helium and iodine vapour. The sensitive region was determined on the one side by the LiF window transmission at 1080 Å and on the other by the photo-ionization threshold of iodine at about 1270 Å, and the quantum efficiency at the Lyman 1216 Å wavelength was claimed as several per cent.

§ 9. Solid State Devices

Although this review is devoted mainly to photometry with photoemissive cells, it seemed good to devote a short and mainly comparative section to some of the photometric devices which depend on internal photoelectric effects.

Photovoltaic or barrier layer cells, in which a thin layer of copper oxide, selenium or lead sulphide is formed on a metal surface and covered with a transparent conducting layer or net, generate a current on illumination. They have long been used for rough photometry, as in photographic exposure meters, and occasionally as comparison detectors for accurate measurement. They are simple in use, and their sensitivity covers the same spectral range as the eye or the usual sort of photographic emulsion, and can be made to agree with either of these by means of filters. Their sensitivity may, however, change with time, they are temperature dependent, and their response varies in a complicated way with the angle of incidence of the light. It is not easy to obtain high sensitivity from a small surface, as would be required in spectrophotometry.

More important in purely scientific work are photoconductive cells which have some importance in the visible spectrum and are the most sensitive detectors available for part of the infra-red.

Cadmium sulphide and selenide are used to make photoresistive detectors for the visible spectrum with a current response of the order of 1 amp per lumen of tungsten light, and in the germanium phototransistor a comparable efficiency is obtained by irradiating a p–n–p double junction. The effective area of this device is, however, small. The most important photoconductive materials used in the infra-red are the sulphide, selenide and telluride of lead, indium antimonide, and germanium doped with small quantities of other elements.

In all these detectors the primary process is the excitation of an electron from one energy level to another; the level separation is typically only a small fraction of an electron volt, so that the response goes well into the infra-red. There is, of course, a certain probability of thermal excitation, so that low-quantum-energy detectors can only be used if cooled. As a rough rule, detectors must be cooled

(conveniently with liquid air) if their response goes beyond 4–5 μ, while the zinc-doped and antimony-doped detectors sensitive to the farther infra-red are cooled to liquid helium temperatures. It is clear that, with cells which are sensitive to the radiation given off copiously by bodies at quite low temperatures, the surroundings of the cell should be cooled as well as the photoconductive layer itself. The essential superiority of the photoconductive devices over thermal detectors lies in the selectivity of the former; the latter are sensitive over an indefinitely wide spectral band and are subject to larger fluctuations which can be calculated thermodynamically (Smith, Jones and Chasmar 1957).

After the initial excitation of an electron has conferred effective mobility on the carriers of charge, conduction can persist until the electron is recaptured. The recapture is not necessarily a simple process with an exponential time constant,

Table 2. Photoconductive and Other Cells

Description	(1)	(2)	Reference
Cs–Sb (photoemissive)	0·4	0·65	
Cs–O–Ag (photoemissive)	0·8	1·2	
CdS (photoconductive)	0·68	0·9	
Germanium phototransistor	1·6	~2	
PbS	2·4	3·5	Smith, Jones & Chasmar 1957
PbSe	3·3	5	Mullard (Catalogue material)
PbTe	4·4	6	
InSb	6·5	8	
Ge crystal doped Au		9·5	
Ge crystal doped Au and Sb		6	Beyen *et al.* 1959
Ge doped Zn		14	
		10–22	Keyes *et al.* 1956
Ge doped Sb	sensitivity low at 50 μ	>118	Fray & Oliver 1959

(1) Wavelength of maximum sensitivity (μ); (2) approximate long-wave limit (μ).

but times ranging from less than a microsecond to hundreds of microseconds are involved. There is therefore an amplification, depending on the time factor and the mobility of the carriers, leading to a high apparent quantum efficiency, though of course the photon noise (§ 1) corresponds to the efficiency of the primary excitation. Photoconductive cells contain, however, other sources of noise which are usually dominant, and which have a frequency distribution rising at low frequencies. The cells are therefore used with ' chopped ' light, and the best chopping frequency depends on the frequency distribution of the noise and the response time of the cell. It is common to use light chopped at a few hundred cycles per second.

The internal behaviour of the doped germanium single-crystal photoconductors is fairly well understood (Beyen *et al.* 1959) but the lead salt detectors develop their full sensitivity, which is usually associated with a fairly long time constant, only after an empirical treatment with oxygen. Their mechanism is more complicated, and the two main explanations involve the existence of electron traps in which electrons are temporarily preserved from recombination or the existence of photosensitive potential barriers at intercrystalline boundaries.

The general characteristics of the known solid-state detectors are given in table 2.

§ 10. Examples of Special Applications of Photoelectric Photometry

The application of photoelectric photometry to stellar problems has been discussed in the section on the photometry of weak light. The laboratory problems which have been attacked are, of course, too numerous to recount, and the examples given below illustrate techniques of special interest or general application.

10.1. *Precision Photometry*

In most photoelectric measurements the consistency achieved over any considerable time interval is probably a few per cent, though rapid comparison methods, and particularly a.c. null methods, allow the photoelectric equality of two light fluxes to be established much more closely. In the maintenance of photometric standards and the applied photometry of light sources, photoelectric cells are used with some precautions to achieve higher accuracy. In the measurements and calculations of illuminating engineering, intensities and fluxes must be weighted according to the spectral sensitivity of the standard (human) observer, and this can be done by filters (Preston 1946). At present caesium–bismuth cells are most used, since their spectral sensitivity is well distributed throughout the visible spectrum. Vacuum cells are employed since their characteristics are stable compared with any type of multiplying cell, but it has been found that even in vacuum cells the emission is influenced by charges on the inner bulb surface (Preston and Gordon-Smith 1955) and cells for precise measurement ought to be designed to avoid this possibility by means of a conducting screen at anode potential completely enclosing the cathode. It is further found that for the highest accuracy the cell should be operated at constant temperature, and it might be desirable to keep the cells, when not in use, in low temperature storage to reduce any migration of caesium within the bulb.

The practice at the National Physical Laboratory is to measure the cell output with a simple balanced circuit of two electrometer valves. The electrical measurement is made into a null one by opposing the potential difference produced by the photocurrent against the e.m.f. of an accurate potentiometer.

An alternative approach to problems of accurate photometry is exemplified by the work of Stair (1951) and others at the National Bureau of Standards. Here the object was spectrophotometry and the method adopted involved chopping the light at 510 c/s, using a vacuum photocell, a selective amplifier and a (non-synchronous) rectifying voltmeter.

10.2. *Multichannel Spectroscopy for Industrial Analysis*

It is obviously possible to arrange for a photocell to scan a spectrum and to make a record of the intensity distribution; commercial instruments for doing this are available, including some in which the photocell is traversed in the focal plane of a spectrograph of established design, and others with the scanning arrangement designed *ad hoc*.

Several firms produce instruments in which a chosen spectral line of each of a number of elements is made to fall on a separate photocell. During exposure to a spark source the radiation in each of the channels is measured and integrated so that quantitative spectrochemical analysis is carried out automatically and with

great speed. The method is particularly valuable in the metallurgical industry, since an analysis can be completed while a charge of molten metal is kept waiting. The first of these instruments was described by Saunderson, Caldecourt and Peterson (1945) and its features illustrate well the technique of photoelectric spectroscopy. The spectrometer is a large grating instrument, and slits are placed at appropriate points on the focal curve. Behind each of these is a photomultiplier of the 1P 21 or 1P 28 type (ultra-violet sensitive). In order to avoid changes in the efficiency of the channels due to movements of the source, each cell is covered with a diffusing screen of ground glass or quartz, and an auxiliary lens focuses the exit slit on each diffusing screen. The current from each multiplier is made directly to charge a high quality (polystyrene) condenser. After the end of the exposure the condensers are allowed to discharge through resistances, and the time taken by each condenser to discharge from its acquired potential to a standard potential is a (logarithmic) measure of the integrated light; these times are automatically recorded and used with calibration curves to give direct analytical results. The whole cycle lasts 40 seconds and the relative *photometric* accuracy is within 1%. The dark current of each cell is separately compensated, and during the exposure each cell is exposed alternately to its line and to the background in an adjacent part of the spectrum, the background being integrated in the condensers with reversed sign.

10.3. *Some Differential Applications*

Since photoelectric detectors produce linearly electrical signals which can readily be amplified and transformed in the mathematical sense, it is sometimes possible to measure with spectacular sensitivity quantities which can be made to appear as the difference of two light fluxes.

An extreme example (in which, however, solid state photocells were actually used) is given by an optical lever (Jones and Richards 1959), made to detect the rotation of a small mirror. An image of an illuminated grid was thrown on another grid after reflection at the mirror. The second grid was formed in two halves displaced by the width of a line, and behind each half was a cadmium selenide photoresistive cell connected to a difference-taking circuit. The sensitivity of this device is limited by mechanical stability and the optical effects of convection currents, but the angular displacement which can be detected (about 10^{-10} radian) is only one millionth of the half width of the diffraction pattern due to the mirror.

An important application of the differential technique is the investigation of the local magnetic fields over the Sun's disc using the Zeeman effect of selected spectral lines (Babcock 1953). An image of the solar disc was thrown on the slit of a large grating spectrograph, and in the focal plane of the instrument a pair of slits is placed on the flanks of a suitable Fraunhofer line. The longitudinal Zeeman effect splits the line into circularly polarized components, but in the small magnetic fields which occur the splitting is small compared with the width of the line. In front of the spectroscope slit is a polarizer preceded by an electro-optical plate (a crystal of ammonium dihydrogen phosphate between transparent electrodes) which under an alternating electric field acquires double refraction and behaves as a quarter-wave plate of alternating sign. Behind each slit is a photomultiplier tube connected to a difference-taking circuit. The alternating output from this

is a measure of the Zeeman effect ; it is amplified, rectified by a vibrator synchronous with the field applied to the crystal plate, and recorded as the Sun's disc is scanned by the spectrograph slit. The random noise of this arrangement is about 0·1 gauss, which corresponds to a Zeeman splitting of less than 10^{-5} Å, which may be compared with a profile width of at least some tenths of an ångström.

A comparatively sophisticated transformation of electrical quantities representing light flux is made in investigating the correlation between the fluctuations of the light intensity at two detectors illuminated by the same source—the so-called intensity interferometry of Hanbury Brown and Twiss (1957). Unlike most photometric measurements this experiment involves critically the time-delay properties of the photomultipliers used, and electron-focusing multipliers are therefore employed (§ 2). The correlation is determined in principle by multiplying together the output currents from the two photocells and integrating over a period ; the complications arise in doing this without introducing any possibility of a drift which would simulate the correlation sought. The principle of the technique adopted is to reverse periodically the phase of one of the input connections and simultaneously to reverse the sense of the final electromechanical integrator. This commutation technique is in fact employed twice, so that it is possible to detect a correlation signal which is very small compared with the fluctuation noise of the system.

Acknowledgments

I am glad to record my indebtedness to Mr. J. S. Preston of the National Physical Laboratory and Mr. J. Sharpe of E.M.I. Electronics for discussion and information used in certain parts of this article.

References

Babcock, H. W., 1953, *Astrophys. J.*, **118**, 387.

Beyen, W., Bratt, P., Davis, H., Johnson, L., Levinstein, H., and MacRae, A., 1959, *J. Opt. Soc. Amer.*, **49**, 686.

Blitzstein, W., 1953, *Astronomical Photoelectric Photometry* (Washington: Amer. Ass. Adv. Sci.) ; 1958, *The Telescope of Moderate Size* (Philadelphia: University of Pennsylvania Press).

Brackmann, R. T., Fite, W. L., and Hagen, K. E., 1958, *Rev. Sci. Instrum.*, **29**, 125.

Breitenberger, E., 1955, *Progr. Nucl. Phys.*, **4**, 56 (London : Pergamon Press).

Chubb, T. A., and Freidman, H., 1955, *Rev. Sci. Instrum.*, **26**, 493.

Clark Jones, R., 1959, *J. Opt. Soc. Amer.*, **49**, 645.

Coblentz, W., 1946, *J. Opt. Soc. Amer.*, **36**, 72.

Colgate, S., 1959, *Rev. Sci. Instrum.*, **30**, 140.

Dunkelman, L., and Lock, C., 1951, *J. Opt. Soc. Amer.*, **41**, 802.

Elster, J., and Geitel, H., 1910, *Phys. Z.*, **11**, 259.

Engstrom, R. W., 1947, *J. Opt. Soc. Amer.*, **37**, 420.

Engstrom, R. W., Staudenheimer, R. G., and Glover, A. M., 1952, *Nucleonics*, **10**, 58.

Fellgett, P., 1955, *Vistas in Astronomy* (London : Pergamon Press) ; 1958 a, *The Telescope of Moderate Size* (Philadelphia : University of Pennsylvania Press) ; 1958 b, *Mon. Not. R. Astr. Soc.*, **118**, 224.

Fray, S. J., and Oliver, J. F. C., 1959, *J. Sci. Instrum.*, **36**, 195.

Garton, W. R. S., Webb, M. S. W., and Wildy, P. C., 1957, *J. Sci. Instrum.*, **34**, 496.

Görlich, P., 1936, *Z. Phys.*, **101**, 335.

Gyorgy, E. M., and Kingston, R. H., 1951, *Phys. Rev.*, **83**, 220.

GYORGY, E. M., PIERCE, E. R., HARVEY, C. G., and KINGSTON, R. H., 1952, *Rev. Sci. Instrum.*, **23**, 8.
HANBURY BROWN, R., and TWISS, R. Q., 1957, *Proc. Roy. Soc.* A, **243**, 291.
HARPER, W. J., and CHOYKE, W. J., 1957, *Rev. Sci. Instrum.*, **27**, 966.
HINTEREGGER, H. E., and WATANABE, K., 1953, *J. Opt. Soc. Amer.*, **43**, 604.
JACQUINOT, P., 1954, *J. Opt. Soc. Amer.*, **44**, 161.
JAMES, J. F., 1959, *J. Sci. Instrum.*, **36**, 188.
JOHNSON, F. S., WATANABE, K., and TOUSEY, R., 1951, *J. Opt. Soc. Amer.*, **41**, 702.
JOHNSON, H. L., 1948, *Astrophys. J.*, **107**, 34.
JONES, R. V., and RICHARDS, J. C. S., 1959, *J. Sci. Instrum.*, **36**, 90.
KEYES, R. J., ZWERDLING, S., FINER, S., and KOHM, H. H., 1956, *Phys. Rev.*, **104**, 1804.
KRON, G. E., 1958, *Publ. Astr. Soc. Pacif.*, **70**, 285.
LALLEMAND, A., 1949, *Le Vide*, **4**, 618; 1953, *Astronomical Photoelectric Photometry* (Washington: Amer. Ass. Adv. Sci.).
LENOUVEL, F., and DAGUILLON, J., 1954, *J. Phys. Radium*, **15**, 287.
MORTON, G. A., 1949, *R.C.A. Rev.*, **10**, 525.
PRESTON, J. S., 1946, *J. Sci. Instrum.*, **23**, 173, 211.
PRESTON, J. S., and GORDON-SMITH, G. W., 1955, *Brit. J. Appl. Phys.*, **6**, 329.
REDMAN, R. O., and YATES, G. G., 1953, *Astronomical Photoelectric Photometry* (Washington: Amer. Ass. Adv. Sci.).
RODDA, S., 1953, *Photoelectric Multipliers* (London: Macdonald).
ROSE, A., 1948, *J. Opt. Soc. Amer.*, **38**, 196.
SAUNDERSON, J. L., CALDECOURT, V. J., and PETERSON, E. W., 1945, *J. Opt. Soc. Amer.*, **35**, 681.
SCHWETZOFF, V., 1954, *Rev. Gén. Élect.*, **63**, 71.
SHARPE, J., 1957, *Brit. Comm. and Electronics*, August issue.
SIMON, H., and SUHRMANN, R., 1958, *Der lichtelektrische Effekt* (Berlin: Springer).
SMITH, R. A., JONES, F. E., and CHASMAR, R. P., 1957, *The Detection and Measurement of Infra-red Radiation* (Oxford: University Press).
SOMMER, W., 1955, *Rev. Sci. Instrum.*, **26**, 725.
STAIR, R., 1951, *J. Res. Nat. Bureau Stands*, **46**, 437.
THORNBURG, W., 1956, *Rev. Sci. Instrum.*, **27**, 99.
WATANABE, K., and INN, E. C. Y., 1953, *J. Opt. Soc. Amer.*, **43**, 32.
YATES, G. G., 1948, *Mon. Not. R. Astr. Soc.*, **108**, 476.

There is an elaborate bibliography of photoelectric methods in *Astronomical Photoelectric Photometry* (1953) published by the American Association for the Advancement of Science, Washington. The Physical Society Reports on "Photometry" (Buckley, H., 1941, *Rep. Progr. Phys.*, **8**, 318) and on the "Infra-red Spectrum" (Sutherland, G. B. B. M., and Lee, H. W., 1947, *Ibid.*, **11**, 150) deal with certain aspects of photoelectric photometry at those dates.

EXPERIMENTAL ANALYSIS OF THE ELECTRONIC STRUCTURE OF METALS

By A. B. PIPPARD
Cavendish Laboratory, Cambridge

CONTENTS

	PAGE
§ 1. Introduction	176
§ 2. The independent-particle model	181
§ 3. Transport phenomena	194
3.1. The method of trajectories	194
3.2. Spatially varying electric fields	196
3.3. Galvanomagnetic effects	198
§ 4. Analysis of special phenomena	205
4.1. Introduction	205
4.2. The de Haas–van Alphen and related effects	206
4.3. The anomalous skin effect	218
4.4. Cyclotron resonance	223
4.5. Ultrasonic attenuation and magneto-acoustic effects	236
4.6. Size effects	255
§ 5. Applications	256
Appendix: The standard metal	261
Acknowledgments	261
References	261

Abstract. After a short résumé of the general ideas and assumptions of the independent-particle model of a metal, an account is given of the experimental methods which have been, or may be, used to determine the details of the model for any given metal, with special reference to the shape of the Fermi surface and the electronic velocity at all points on the Fermi surface. Particular attention is paid to the exposition of the theory underlying each method, and as far as possible only simple mathematical and physical ideas are used. The conditions of application of the methods are discussed, and at the end examples are given of the results so far obtained by their use, with special emphasis on the analysis of the electronic structure of copper. The methods discussed are the following: Magneto-resistance, de Haas–van Alphen and Schubnikov effects, anomalous skin effect, cyclotron resonance, ultrasonic attenuation and magneto-acoustic effects, size effects.

§ 1. Introduction

THE subject of this review is the use of experiment in obtaining as accurate as possible a description of a real metal, so that its properties, particularly those concerned with the transport of electrical charge and of thermal energy, may ultimately be understood more exactly than at present. It should, of course, be appreciated from the outset that no strictly exact description is possible. Even under the most ideal conditions, with electrons moving in a perfectly ordered ionic lattice at the absolute zero, exact specification involves writing down the ground state wave function of the assembly of electrons and nuclei—an impossible task. But fortunately this is not what is required. If there were any serious doubt about whether in principle the laws of quantum mechanics are adequate to account for the observed properties of metals one might have to consider questions of this sort. The task, however, is not to see

whether the fundamental laws give a correct description, but to discover a mental technique which will avoid as far as possible the necessity of referring to fundamental laws—to find a model whose properties can be apprehended instinctively, so that any calculations of these properties can be seen in their proper perspective, as more precise statements of what the physical imagination is prepared to accept as qualitatively intelligible. One cannot hope that the model will reproduce every detail of the experimental truth, but one may aim at a reasonable compromise ; a model not so elaborate as to bemuse the imagination, yet close enough to the truth to act as an acceptable substitute for it, above all one whose limits of validity are understood. Just how good a model one may expect to devise is a matter for speculation—certainly nothing like a satisfactory solution has been found yet, as can be seen from the tables comparing theoretical prediction with experimental results, to be found in any standard text on the theory of metals. Yet the limitations of existing models may not lie so much in the basic assumptions as in the introduction of simplifying approximations to make them analytically tractable. There is, in fact, a real need for the analysis, by numerical methods if all else fails, of models more generally formulated than some which have been studied in the past. But if the lack of such analysis is to be imputed a fault on the part of the theoreticians, the experimenters for their part cannot escape blame ; for in spite of the huge mass of information on almost every conceivable metallic property, it does not appear to be sufficient to allow an adequate model of a metal to be synthesized from it. The history of the development of metal physics has been one of invention, on theoretical grounds, of successively more plausible models, whose success has been gauged by comparison with experiment ; rarely has experiment led directly to an improvement in the model. The reason for this is to be found in the complexity of the problem, the properties which are most easily studied being determined by a large number of different parameters of the model, so that it is hard to make any definite statement about the values of these parameters by an examination of the data. In the last few years the situation has improved somewhat by the discovery of phenomena which are dependent on only a few parameters, and there is now some hope that a model of a metal may be synthesized on the basis of experiment and some rather general theoretical ideas. It is these phenomena with which we shall be concerned.

Underlying the whole treatment that follows is one great simplifying assumption, which certainly must lead to errors, but which it may be hoped will not be disastrous. This assumption is that the dynamics of the conduction electrons in the metal may be reduced to the quantum dynamics of individual particles, as if there were no interactions between the electrons. We shall not discuss in detail the measure of *a priori* justification which can be given to this assumption, but shall indicate briefly why it may be expected to work fairly well. In this discussion, as in the whole of what follows, we shall often adopt another approximation which, though perhaps unnecessary, may be regarded as a simplification of thought, especially by experimental physicists ; as far as possible we shall avoid the formalism of quantum mechanics and adhere to the representation of semi-classical mechanics, such as is associated with the names of Bohr, Sommerfeld and Wilson ; that is, the conceptual framework of classical mechanics patched up as necessary by empirical rules of quantization. It is desirable to emphasize that

this attitude is envisaged as a convenience and an aid to the imagination; it is only adopted when it is believed that the errors involved are negligible.

The simplest model of a metal is the free-electron model; the electrons are treated as if they formed a perfect gas of non-interacting particles moving in a rectangular box in which the potential is uniform. Since electrons are fermions the many-particle wave function is a determinant whose elements are individual plane wave functions, each of which is interpreted semi-classically as an individual unlocalized particle having well-defined momentum. Alternatively, by forming linear combinations of the plane waves, wave packets may be constructed to represent partially localized electrons whose momentum is correspondingly uncertain. Now in reality the electrons interact with one another through the long-range Coulomb potential e^2/r. Even although the mean electronic charge is neutralized by the positive lattice charge, local fluctuations in electron density set up fields which influence the whole assembly. So strong is this effect, indeed, that attempts to allow for it by perturbation methods fail through divergence of the distant interaction. A more powerful method of attack, due principally to Bohm and Pines (Pines 1955), analyses the fluctuations of density into their Fourier components, introducing collective variables for the electron assembly in a manner analogous to Debye's use of lattice vibration coordinates to describe the motion of the strongly interacting atoms of a solid. Each Fourier component of charge fluctuation (*plasma mode*) is a fairly independent entity with a characteristic frequency of oscillation $\omega_p = (4\pi Ne^2/m)^{1/2}$, where e is the electronic charge in e.s.u. To a first approximation all plasma modes have the same frequency. The quantum energy associated with a plasma oscillation, $\hbar\omega_p$, is many volts for most metals, so that at normal temperatures these oscillations are unexcited. If then we neglect the zero point plasma oscillations as of no account, we are led to believe that the spontaneous motions of the electrons are so linked as to set up no fluctuations in charge density. This is, however, true only as concerns fluctuations over distances rather larger than the mean distance between electrons, since there is a lower limit to the permissible wavelength for a plasma oscillation. In effect, then, the motions of electrons can be thought of as essentially uncorrelated, except in so far as they arrange to screen the long-range part of the Coulomb interaction, that is, their motions are correlated to the extent that no long-range variations of charge density occur, and in effect their interaction potential instead of being e^2/r approximates to the form $(e^2/r)\,e^{-\kappa r}$, vanishing rapidly at large distances. The total number of permitted plasma modes is rather small compared with the number of electrons, and this means that the number of degrees of freedom described by independent collective modes is a small proportion of the total, and the restriction to freedom of movement of the particles fairly slight. In this respect there is a marked contrast to the Debye solid, in which all motions can be described in terms of independent oscillations, and the strongly interacting atoms of the solid may be replaced by a gas of phonons, quasi-particles whose motions are practically independent, but whose character is totally different from that of the atoms composing the solid. But in the metal the quasi-particles may be considered to resemble electrons very closely. It is probably not significant to enquire precisely how many quasi-particles there are, for, being fermions, they are nearly all packed down into the lowest-lying energy levels, one to each level in accord with the

Pauli exclusion principle, and the only ones that take part in transport and related properties are those lying within a few kT of the highest occupied energy level, the *Fermi level*. These properties are in fact determined by the wave number $\mathbf{k}$ of a quasi-particle at the Fermi level, and by the relation between energy and wave number for quasi-particles excited above the Fermi level and for holes left below the Fermi level by such excitations. There is no real need in discussing these properties to consider in any detail the precise description of the ground state of the assembly; it is enough to know that the excitations are particle-like, and to know the relation between energy and wave number and the interactions between them. Since, however, they are particle-like it may be convenient to relate the quasi-particle picture to the model of non-interacting electrons, as we have just done by imagining the ground state to consist of an assembly of independent particles packed into the lowest states, the relation between ϵ and $\mathbf{k}$ at the Fermi level being chosen to correspond to the properties of the excitations. To take the quasi-particles as entirely independent is perhaps to go a little too far, since the screened Coulomb interaction is by no means obviously negligible. We must discuss what effect it is likely to have on the behaviour of the quasi-particles, in order to show that it is probably only a minor influence.

The intervention of the exclusion principle is of paramount importance in restricting the effect of electron–electron interactions. If the electrons (as we shall now refer to the quasi-particles) were not fermions and had a cross section for collision determined by the screened Coulomb potential, the mean free path between collisions of an electron in a metal would be only a few ångströms, the lifetime being correspondingly about 10^{-16} seconds. According to the uncertainty principle the energy of the single-particle state would be indeterminate to the extent of about 10 volts, and we should be unable to think of the electrons as even approximately independent unless we were considering phenomena involving energy interchanges of at least 10 volts. When we take account of the exclusion principle, however, everything is changed. In the collision of two electrons energy and momentum are conserved, and the states into which the electrons are scattered must both be unoccupied—in effect they must be states near or above the Fermi level. Thus if energy is to be conserved the only electrons which are already present and can collide are those lying in the thermally excited region near the Fermi surface. As a result any one electron belonging to this region cannot collide with each of the N electrons present, but only with something like NkT/ϵ_0, where ϵ_0 is the Fermi energy. Moreover, even if the two electrons come from the permitted region, not all collisions which conserve energy and momentum are permitted. For if one electron gains more than a few kT of energy the other will find itself with that much less, and no unoccupied state in which to fall. This requirement of no energy exchange exceeding a few kT imposes another factor of the order kT/ϵ_0 on the scattering probability. As a result the mean free path is not a few ångströms but something like $(\epsilon_0/kT)^2$ ångströms, say 10^{-4} cm at room temperature, 10 cm at 1°K. The energy uncertainty corresponding to this free path is smaller than the thermal spread of electron energies by a factor of about T/T_0, where T_0 is the degeneracy temperature ϵ_0/k, say 50 000°K. There is thus no serious bar at ordinary temperatures to treating the particles as independent for the purpose of thermodynamical

calculations. Moreover, at no ordinary temperatures does the probability of electron–electron collision approach that of electron–phonon or, at the lowest temperatures, electron–impurity collision. In discussing transport phenomena we may therefore regard the electrons as virtually independent particles suffering collision with thermally excited phonons or with static imperfections of the lattice. There is experimental confirmation of this view in comparisons between electrical conductivity σ and thermal conductivity κ of the same samples. If the dominant mechanism of electron scattering is elastic collisions with static imperfections, as at very low temperatures, the Wiedemann–Franz law is expected to hold under rather general conditions, the ratio $\kappa/\sigma T$ taking the value $\frac{1}{3}\pi^2 k^2/e^2$. Now the destruction of an electric current involves the transfer of momentum to the lattice, and electron–electron collisions do nothing to aid this, while conversely the destruction of a thermal current only involves a rearrangement of energy among the electrons, such as can be effected by electron–electron collisions as well as by electron–lattice collisions. The experimental verification of the Wiedemann–Franz law, even in extremely pure metals, with the correct numerical value of the ratio, shows that the probability of electron–electron collision is much less than that of electron–lattice collision.

By arguments of this sort, which have been given detailed treatment only for the over-idealized free electron model, we may support the conviction that it is worth while adopting the independent-particle hypothesis in discussing the electronic structure of real metals, in which the assumption of a uniform potential is certainly inadequate. It may be that experiments will ultimately show that no consistent independent-particle model can be devised, but this has not yet happened. It seems likely that, for most of the phenomena which we shall treat, the discrepancies attributable to electron–electron interactions will be quite small. If this should turn out to be a false hope, the model will have failed and more elaborate models will need to be devised ; there is little point in attempting to meet this situation yet.

Finally, one more simplifying assumption must be mentioned, the neglect of electron–lattice interactions of the type now known to be responsible for superconductivity. Although it seems certain that these interactions are strong enough in many metals to reverse the sign of the short-range interelectronic forces between electrons close to the Fermi energy, and to lead to a collective interaction at low enough temperatures, we shall assume that so long as a metal is not superconducting the independent-particle picture is still adequate, though it must be remembered that the electron–lattice interaction may play its part in determining the dynamical properties of electrons near the Fermi level. One might expect that all effects due to this interaction would show some dependence on M, the isotopic mass of the nucleus—for example, the superconducting transition temperature varies as $M^{-1/2}$—and the small amount of evidence that the electronic specific heat in the normal (non-superconducting) state is independent of the isotopic mass gives support to our neglect of the interaction. Present theories of superconductivity assume that in the normal state an independent-particle model is adequate, and that in the superconducting state the excitations have something of the character of the single particles in the normal state. If the present feeling of optimism about these theories should continue to be justified

by more detailed experimental confirmation, perhaps this will be the surest support for the independent-particle model of a normal metal, for all types of electron–electron and electron–lattice interactions contribute in an important measure to the structure of the superconducting state.

§2. The Independent-particle Model

The independent-particle model in its primitive form, as initiated by Sommerfeld and Bloch, disregards interactions between the electrons, except in so far as the mean Coulomb field may contribute to the potential in which each electron is moving independently. Under this assumption the problem of determining the properties of the metal is well defined, in the sense that a specification of the periodic potential determines the permitted energy states for the electrons, and from this all the properties follow in principle. We shall not concern ourselves here with an account of the elaborate techniques which have been devised to carry out this programme in specific instances (Herman 1958, Reitz 1955, Slater 1956), but shall confine our attention to those aspects of the problem which are common to all models of this type, and which are particularly relevant to the special topics under review.

We consider a particle moving in a periodic potential $V(\mathbf{r})$, having the property that $V(\mathbf{r}+l\mathbf{a}+m\mathbf{b}+n\mathbf{c}) = V(\mathbf{r})$, where l, m, n are integers, and $\mathbf{a}, \mathbf{b}, \mathbf{c}$ vectors defining the parallelepipedal unit cell which, replicated through all space, generates the periodic potential. The wave equation for the particle in such a potential has solutions which are analogous to the plane wave solutions for a particle moving in a uniform potential. The wave function takes the same form in each cell, but the phase of the function varies from cell to cell in a linear manner with position of the cell. Thus if $\mathbf{R}$ represents a point in a cell, measured from an origin in the cell, the wave function may be written

$$\psi(\mathbf{r}) = U(\mathbf{R})\, e^{-i\mathbf{k}.(l\mathbf{a}+m\mathbf{b}+n\mathbf{c})}$$

where

$$\mathbf{r} = \mathbf{R}+l\mathbf{a}+m\mathbf{b}+n\mathbf{c}. \qquad \ldots\ldots(1)$$

Clearly by the substitution $U(\mathbf{R})\, e^{i\mathbf{k}.\mathbf{R}} \equiv u(\mathbf{R})$, this can be re-expressed in a way which accentuates the similarity to a plane wave

$$\psi(\mathbf{r}) = u(\mathbf{R})\, e^{-i\mathbf{k}.\mathbf{r}}. \qquad \ldots\ldots(2)$$

If $V(\mathbf{r})$ were constant, all solutions would have $u(\mathbf{R}) =$ constant; the nature of the periodic potential reveals itself in the form of $u(\mathbf{R})$ for any given $\mathbf{k}$, as well as in the energy of the particle whose wave function is characterized by $\mathbf{k}$.

The representation of $\psi(\mathbf{r})$ as a function $u(\mathbf{R})$ and a phase factor $e^{-i\mathbf{k}.\mathbf{r}}$, as in (2), is not unique; for we may write

$$\psi(\mathbf{r}) = u(\mathbf{R})\, e^{-i\mathbf{k}.\mathbf{r}} = u'(\mathbf{R})\, e^{-i(\mathbf{k}+\mathbf{K}).\mathbf{r}},$$

provided that $e^{-i\mathbf{K}.\mathbf{r}}$ is a function of $\mathbf{R}$ alone, so that $u'(\mathbf{R})\, e^{-i\mathbf{K}.\mathbf{r}} = u(\mathbf{R})$. This implies, from (1), that

$$e^{-i\mathbf{K}.(l\mathbf{a}+m\mathbf{b}+n\mathbf{c})} = \text{constant}. \qquad \ldots\ldots(3)$$

The constant can only be unity, and the vectors $\mathbf{K}$ which satisfy the equation form the *reciprocal lattice*. To construct this, consider the three vectors $\mathbf{K}_a$, $\mathbf{K}_b$ and $\mathbf{K}_c$ having the properties

$$\mathbf{K}_a.\mathbf{a} = 2\pi, \quad \mathbf{K}_a.\mathbf{b} = \mathbf{K}_a.\mathbf{c} = 0, \qquad \ldots\ldots(4)$$

and similarly for the other two. Then it is clear that if $\mathbf{K}$ is a sum of integral multiples of $\mathbf{K}_a$, $\mathbf{K}_b$ and $\mathbf{K}_c$ the exponent in (3) is an integral multiple of $2\pi i$, and (3) is satisfied.

As an example of this construction consider a crystal structure which is common in metals, the face-centred cubic lattice of figure 1. This may be derived from a unit cell containing one atom, whose vectors are $a_0(-1, 1, 0)$, $a_0(-1, -1, 0)$ and $a_0(1, 0, 1)$, where $2a_0$ is the cube side. It is easy to see that

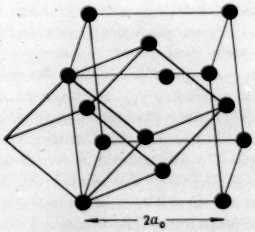


Figure 1. The face-centred cubic lattice, showing unit cell containing one lattice point.

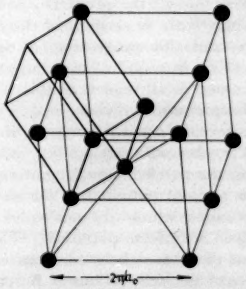


Figure 2. The body-centred cubic lattice, showing unit cell containing one lattice point.

the reciprocal lattice vectors which satisfy (4) are $(\pi/a_0)(-1, 1, 1)$, $(\pi/a_0)(-1, -1, 1)$ and $(\pi/a_0)(0, 0, 2)$; these define a cell which on repetition generates the body-centred cubic lattice, as in figure 2. Any vector $\mathbf{K}$ joining two points of this lattice may be added to $\mathbf{k}$ in (2) to give (with suitable modification of $u(\mathbf{R})$) an equivalent expression for $\psi(\mathbf{r})$. Thus the energy of an electron may be represented as a periodic function of $\mathbf{k}$; if $\epsilon(\mathbf{k})$ is known within one unit cell of the reciprocal lattice, it is known for all $\mathbf{k}$. As a matter of convenience it is usual to redefine the unit cell so as to exhibit more clearly its cubic symmetry; any dissection of the cell in figure 2, with reconstruction after translation of some of its parts through a reciprocal lattice vector, is a permissible redefinition, and the cell of figure 4 may be considered to arise in this way. It contains one lattice point as its centre, and by repetition fills all space and generates the body-centred cubic lattice. An equivalent construction for the face-centred cubic lattice is shown in figure 3. If the metallic lattice is face-centred cubic, figure 3 is the *atomic polyhedron* and figure 4 the *first Brillouin zone* in reciprocal space; while for a body-centred cubic lattice the diagrams are reversed, the atomic polyhedron having the form of figure 4 and the Brillouin zone the form of figure 3.

We have seen that a specification of $\epsilon(\mathbf{k})$ within the Brillouin zone provides complete information about the solution of the wave equation, but it must be realized that for a given potential $V(\mathbf{r})$ there is an infinite number of solutions corresponding to each $\mathbf{k}$, having different energies and different $u(\mathbf{R})$. This is

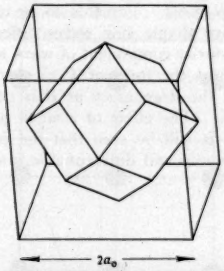


Figure 3. The atomic polyhedron for a face-centred cubic lattice.

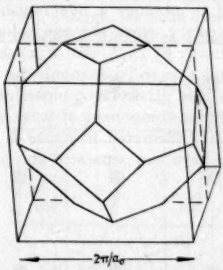


Figure 4. The first Brillouin zone for a face-centred cubic lattice.

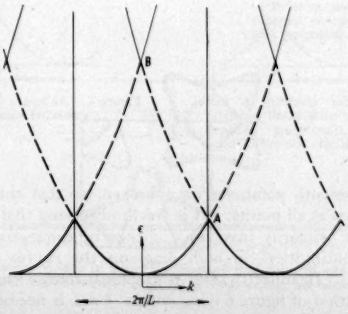


Figure 5. $\epsilon(k)$ for free electron in one dimension, with artificial zone structure imposed.

simply illustrated by an almost trivial one-dimensional example, a particle moving on a line in a uniform potential so that $\epsilon = \hbar^2 k^2/2m$. Let us now divide the line into segments of length L, thus artificially defining a cellular structure ; a typical solution of the wave equation may now be written

$$\psi(x) = e^{-ikx} = e^{2\pi i X/L}\, e^{-i(k+2\pi/L)\,x} \quad \text{etc.},$$

in which the variation of $\psi(x)$ in a single cell, $u(X)$, takes the successive forms $1, e^{2\pi i X/L}, \ldots, e^{2\pi n i X/L}, \ldots$, and the value of k is correspondingly changed by multiples of the reciprocal lattice vector $2\pi/L$. The curves of $\epsilon(k)$ are shown in figure 5. The different parabolae correspond to different choices of K $(= 2\pi n/L)$,

and the thickened and broken curves are the lowest two branches of the $\epsilon(k)$ relation in the periodic k-space.

Let us now see what happens when $V(x)$ is not constant, but varies periodically with period L. Then the two degenerate solutions at the point A may have their degeneracy removed if their wave functions combine with the periodic potential to produce a non-vanishing matrix element. For this to be so it is only necessary, since the two solutions at A are simple sine waves differing in wave number by $2\pi/L$, that $V(x)$ shall have a Fourier component of wave number $2\pi/L$, and this will commonly be the case unless the form of $V(x)$ specifically excludes this particular component. Similarly the degeneracy at B is removed by a Fourier component of wave number $4\pi/L$. The effect of a small periodic potential is illustrated in figure 6, from which it will be seen that the periodic branches have now separated and become continuous and differentiable functions,

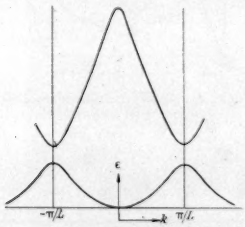


Figure 6. $\epsilon(k)$ for electron in one dimension, with small periodic potential, illustrating reduced and periodically extended zone schemes.

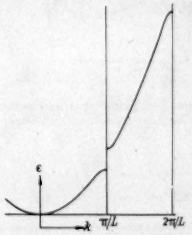


Figure 7. As figure 6, but displayed in extended zone scheme.

while the original smooth parabolas have broken, so that there are energy gaps between the branches at all points. It is worth observing that the transition from figure 5 to figure 6 primarily involves a change of connectivity rather than the appearance of discontinuities—in both diagrams the curves are analytic but of quite different form. In order to give in graphical form a complete specification of $\epsilon(k)$, only the portion of figure 6 lying within $\pm\pi/L$ is necessary; this is known as the *reduced zone scheme*. Alternatively, it may be convenient, especially if the $\epsilon(k)$ curve can thereby be made to approximate to a parabola, to allow k to run between $\pm\infty$ and to display $\epsilon(k)$ as a single-valued function, as in figure 7; this is the *extended zone scheme*. Finally, it will be useful to bear in mind that the reduced zone scheme may be repeated at will, as is begun in the lower curve of figure 6, to produce a multi-valued function $\epsilon(k)$ extending over all values of k; this is the *periodically extended zone scheme*.

The change in connectivity of the $\epsilon(k)$ curves, resulting from the removal of degeneracies at the edges of the Brillouin zone, can make considerable differences to the form of the lines and surfaces of constant energy when we treat the motion of a particle in two- and three-dimensional lattices. As an example, consider a square two-dimensional lattice as in figure 8, whose reciprocal lattice is the square

lattice of figure 9. It is not difficult to see that, in general, any reciprocal lattice vector **K** represents a possible component present in the Fourier analysis of the corresponding lattice potential $V(\mathbf{r})$ into components of the type $e^{-i\mathbf{K}.\mathbf{r}}$. For it is necessary, if this component is to be present, that $(1/N)\Sigma e^{-i\mathbf{K}.(la+mb+nc)}$, summed over all N lattice points, shall not vanish; this is clearly ensured by the definition of the reciprocal lattice. It should be mentioned, though we shall pursue this matter no further, that this is not a sufficient condition for the occurrence of a

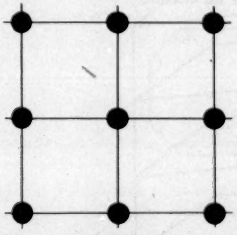


Figure 8. Two-dimensional square lattice.

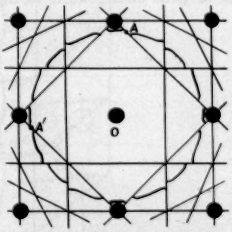


Figure 9. Reciprocal lattice of figure 8, showing construction of zone boundaries, and a typical line of constant energy (for small periodic potential).

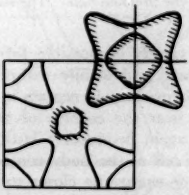


Figure 10. Lines of constant energy in figure 9 remapped into reduced and (top right-hand corner) periodically extended zone scheme. Regions of lower energy are indicated by shading.

given Fourier component. For the present example we assume that all Fourier components defined by the reciprocal lattice are present.

Now, taking one reciprocal lattice point as origin, draw the perpendicular bisectors of all lines joining it to other reciprocal lattice points, as in figure 9. Then, from the construction, any point A on one of these lines (which are the boundaries of higher-order Brillouin zones) may be associated with a point A′ such that AA′ is a reciprocal lattice vector and $OA = OA'$. Thus electrons whose wave vectors are $\overrightarrow{OA}$ and $\overrightarrow{OA'}$ have the same energy if $V(\mathbf{r})$ is constant, and are capable of having their degeneracy removed by the lattice potential when this is taken into account. If the lattice potential is only very small the separation of originally degenerate levels will also be very small, so that a line of constant energy, originally a circle, is only slightly deformed as it crosses the Brillouin

zone boundaries. The new line of constant energy may be remapped into the first Brillouin zone to give a reduced zone picture, as shown in figure 10, of which one corner shows the beginning of the periodic extension of the diagram. In this way it will be seen how the removal of degeneracy breaks up the line of constant energy into three separate branches, which are to be regarded as contours on three distinct sheets of the function $\epsilon(\mathbf{k})$. Moreover, as the shadings reveal, in two of these branches the regions of lower energy lie within the closed curves, as in the original circle, while in the third they lie outside.

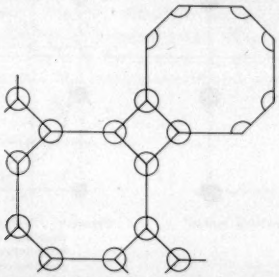


Figure 11. Section of periodically extended zone scheme for face-centred cubic lattice, showing energy surface near corners of the Brillouin zone. The section shown is indicated by a broken line in figure 4.

In three dimensions, the variety of possible forms of energy surface is even greater. If the surface is closed and wholly within the first Brillouin zone, the periodically extended zone scheme merely repeats it as a lattice of isolated closed surfaces. If the surface lies near the corners of the first Brillouin zone, as in figure 11, periodic extension again produces a lattice of isolated closed surfaces (in this case six to every unit cell of the body-centred reciprocal lattice); and if in figure 11 the states of lower energy lie closer to the centre of the zone, then each little closed surface in the extended scheme encloses states of higher energy. A third possibility, which is particularly to be expected with lattices having a high axial ratio, is shown in figure 12, the extended zone scheme giving an array of tubes. And lastly, a multiply-connected surface may arise from contacts with the zone boundary; a simple example may be seen in figure 14. Just as in figures 9 and 10 a very small lattice perturbation of the circular energy curve could result in quite a complex picture in the extended zone scheme, so a repetition of the construction for a three-dimensional lattice gives a remarkable variety of constant energy surfaces. A particular case has been discussed by Gold (1958), the remapping (with rounding-off of corners) of the sphere which in figure 4 would have a volume just equal to the cube, and which therefore lies entirely outside the first Brillouin zone. One of the resulting surfaces in the extended scheme is shown in figure 13.

In the preceding discussion we have assumed that ϵ is a continuous function of $\mathbf{k}$, and indeed for most purposes this is a reasonable assumption. Strictly,

however, in a finite sample, only discrete values of **k** are permitted, such as will allow the wave function to satisfy boundary conditions at the sides of the sample. If we employ the conventional periodic boundary conditions, and a rectangular sample, the permitted values of **k** are such that an integral number of wavelengths

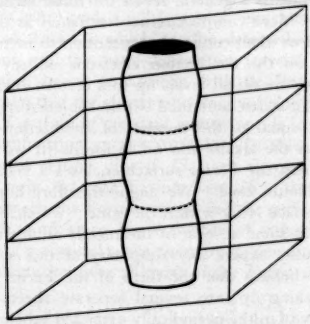


Figure 12. Tubular energy surface in periodically extended zone scheme.

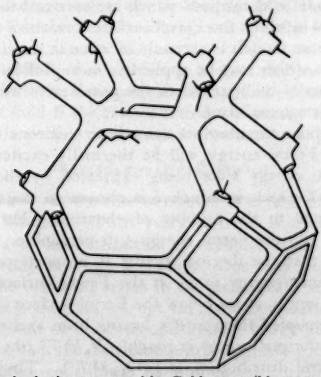


Figure 13. Tubular lattice, suggested by Gold as possible energy surface for lead. (*After* Gold 1958.)

will fit into the sample. For a sample of sides A, B and C, we must have $k_x A = 2n_x\pi$, $k_y B = 2n_y\pi$, $k_z C = 2n_z\pi$, and this implies that in reciprocal space the permitted values of **k** form a rectangular lattice, whose unit cell has volume $8\pi^3/V$, where $V = ABC$, the volume of the sample. The exclusion principle allows each of the states represented by a lattice point to be occupied by not more than two electrons (one of each spin), so that the maximum permitted density of

occupation is $V/4\pi^3$ electrons per unit volume of reciprocal space. From figure 4 it will be seen that the volume of the Brillouin zone for a face-centred cubic lattice is $4\pi^3/a_0^3$, so that the zone will hold V/a_0^3 electrons. Now in a volume V, from figure 3, there are $\frac{1}{2}V/a_0^3$ atoms; therefore the Brillouin zone will hold two electrons per atom. This is a general result for those structures whose unit cell holds only one atom. More complex structures (such as that of tin) need more careful discussion, but we shall confine our attention to metals of simpler form.

At the absolute zero the conduction electrons occupy the lowest available energy states, and therefore fill all levels up to a certain energy, the *Fermi energy*, such that the number of states contained within the corresponding energy surface (*Fermi surface*) is just equal to the number of conduction electrons. For one-electron atoms, such as the alkalis and noble metals (if we assume that the core electrons can be ignored), the Fermi surface encloses a volume equal to one-half the volume of the Brillouin zone. We might therefore hope that it would be a fairly simple closed surface lying within the zone; we shall see that this hope is probably not realized in some at least of the metals of this group. For no other metals can we reasonably expect any simplicity of this sort, and our previous discussion leads us to believe that the form of the Fermi surface may be very strange, probably breaking up into several separate sheets of which some may have complex connectivity in the periodically extended scheme. It is this expectation which makes doubtful the value of attempts to discuss the properties of real metals in terms of model $\epsilon(\mathbf{k})$ relations which are easy to handle analytically, such as the two-band model in which the Fermi surface consists of two sets of ellipsoids. Detailed analysis of these models is certainly of value in developing and comparing mathematical methods which may be applied to more difficult situations, but very little meaning can usually be attached to the parameters which are chosen to fit a particular set of experimental data.

At temperatures above the absolute zero those electrons with energy within a few times kT of the Fermi energy will be thermally excited, the probability of occupation of a given energy state being expressed by the Fermi distribution function $\{\exp(\epsilon-\zeta)/kT+1\}^{-1}$, in which ζ is chosen so that the total number of occupied states is equal to the number of electrons. For most purposes the distinction between ζ and the Fermi energy ϵ_0 is negligible. It is clear from this distribution function that the thermal energy is proportional to the number of electronic states per unit energy range at the Fermi surface; if this is written as N_0, the number of empty states below the Fermi surface is of the order $N_0 kT$, the electrons which occupied them at 0°K having been excited by energies of the order kT, so that the thermal energy is roughly $N_0 k^2 T^2$ (the true value, obtained by integrating the Fermi distribution, is $\frac{1}{3}\pi^2 N_0 k^2 T^2$). Thus the specific heat of the electrons is proportional to the temperature, $C=\gamma T$, the constant γ being $\frac{2}{3}\pi^2 N_0 k^2$. Measurements of γ, for which reliable data now exist for many metals, enable N_0 to be computed. In the independent-electron model N_0 may be related to the velocities of electrons on the Fermi surface, in the way we shall now discuss.

We have so far concerned ourselves only with individual electronic states, represented by plane waves of definite wave number $\mathbf{k}$ and frequency ω $(=\epsilon/\hbar)$. The picture can be made to resemble a classical particle-picture by constructing wave-packets out of a group of plane waves centred on a particular value of $\mathbf{k}$.

This group moves through the metal with a group-velocity $\text{grad}_k\,\omega$, e.g. the x-component of the velocity is $(1/\hbar)\partial\epsilon/\partial k_x$.† The motion of the group determines the motion of the electron which it describes, so that it is permissible to say that an electron described by a wave vector $\mathbf{k}$ is travelling through the metal with velocity $\mathbf{v}(\mathbf{k}) = \hbar^{-1}\,\text{grad}_k\,\epsilon$, contributing to the current an amount $e\hbar^{-1}\,\text{grad}_k\,\epsilon$ and to the momentum $m\hbar^{-1}\,\text{grad}_k\,\epsilon$. In general, one must distinguish between the real momentum $m\hbar^{-1}\,\text{grad}_k\,\epsilon$ which can be detected in principle by macroscopic ballistic methods, and the quasi-momentum (crystal momentum) $\hbar\mathbf{k}$ which is a formal construction based on analogy with de Broglie's relation for free particles. It will be seen that the electronic velocity, being proportional to $\text{grad}_k\,\epsilon$, is directed normal to a surface of constant energy and is inversely proportional to the separation of neighbouring energy surfaces. In this way it is immediately related to the density of states. For if dS is an element of area of the Fermi surface, construct a right cylinder having base dS; then the volume of this cylinder contained between the Fermi surface and the surface of energy $d\epsilon$ greater will be $dS/|\text{grad}_k\,\epsilon|\,d\epsilon$, and since for a sample of volume V there are $V/4\pi^3$ states per unit volume of reciprocal space, the cylinder contains $VdSd\epsilon/[4\pi^3|\text{grad}_k\,\epsilon|]$ states. Hence, by integration over all parts of the Fermi surface, we have that

$$\frac{N_0}{V} = \frac{1}{4\pi^3}\int\frac{dS}{|\text{grad}_k\,\epsilon|} = \frac{S}{4\pi^3\hbar}\left(\overline{\frac{1}{v_0}}\right), \qquad \ldots\ldots(5)$$

where $\overline{(1/v_0)}$ is the average value of the reciprocal of the *Fermi velocity* (the velocity of an electron at the Fermi surface), and S is the total area of the Fermi surface.

The wave-packet picture of an electron in motion enables us to see immediately, by classical analysis, how the electron reacts to applied electric and magnetic fields. In a uniform electric field $\mathbf{E}$ the rate of transfer of energy from the source of the field to the electron is $e\mathbf{v}.\mathbf{E}$; since $\dot{\epsilon}$ can be written as $\dot{\mathbf{k}}.\text{grad}_k\,\epsilon$, it follows that

$$\dot{\mathbf{k}}.\text{grad}_k\,\epsilon = (e/\hbar)\,\mathbf{E}.\text{grad}_k\,\epsilon,$$

which is satisfied if $\hbar\dot{\mathbf{k}} = e\mathbf{E}$, the rate of change of crystal momentum being equal to the force on the electron. Thus every electron is affected in the same way by an electric field, and the distribution of occupied states in reciprocal space is simply shifted uniformly at a rate $\dot{\mathbf{k}} = e\mathbf{E}/\hbar$. One may allow for zone boundary effects by plotting the occupied states in the periodically extended zone scheme, and allowing the whole diagram, except the zone boundaries, to shift uniformly, while only taking account of those occupied states which lie within the first zone. From this it is obvious that the distribution in a completely filled zone is unchanged by a field, and therefore contributes nothing to the current. If the zone is not filled it is easy to write down an expression for the rate of increase of current density under the influence of a uniform field. Consider a small area $d\mathbf{S}$ of the

† To demonstrate this we observe that the phase of a plane wave at $(t, \mathbf{r})$ is given by the expression $\phi = \omega t - \mathbf{k}.\mathbf{r}$. Now any recognizable feature of the wave group owes its character to the particular phase-relationship existing between the various plane-wave components, and if we move through the lattice at such a velocity $\mathbf{v}$ that this relationship is maintained we shall keep in step with the group. To effect this $\mathbf{v}$ must be such that when $\mathbf{r} = \mathbf{v}t$, $\text{grad}_k\,\phi$ does not change with time, i.e.

$$\text{grad}_k\,(\omega - \mathbf{k}.\mathbf{v}) = 0.$$

Since $\text{grad}_k\,(\mathbf{k}.\mathbf{v}) = \mathbf{v}$, $\mathbf{v} = \text{grad}_k\,\omega$.

Fermi surface, imagined, as at 0°K, to be a sharp division between occupied and unoccupied states. After time dt the boundary of occupation has shifted by $e\mathbf{E}dt/\hbar$, so that a new element of occupied volume $edt\mathbf{E}.d\mathbf{S}/\hbar$ has appeared, containing $edt\mathbf{E}.d\mathbf{S}/(4\pi^3\hbar)$ electrons per unit volume of metal. Since each electron moves with velocity $\hbar^{-1}\operatorname{grad}_k \epsilon$, the contribution of this element to the current density is $e^2\,dt\mathbf{E}.d\mathbf{S}\operatorname{grad}_k \epsilon/(4\pi^3\hbar^2)$, so that we may write for the rate of increase of the total current density

$$\mathbf{j} = \frac{e^2}{4\pi^3\hbar^2}\int(\operatorname{grad}_k \epsilon)\,\mathbf{E}.d\mathbf{S}, \qquad \ldots\ldots(6)$$

the integral being taken over the whole Fermi surface. For a crystal of cubic symmetry this expression must be independent of the orientation of $\mathbf{E}$, and is therefore unchanged by averaging over all orientations of $\mathbf{E}$ relative to the crystal axes. From this averaging it follows immediately that

$$\mathbf{j} = \frac{e^2\mathbf{E}}{12\pi^3\hbar^2}\int|\operatorname{grad}_k \epsilon|\,dS = \frac{e^2\,\mathbf{E}S\overline{v_0}}{12\pi^3\hbar}. \qquad \ldots\ldots(7)$$

This expression may also be used with reasonable accuracy for a polycrystalline sample of a non-cubic metal. It may be observed that for a quasi-free-electron model of a metal, having N electrons per unit volume, and with ϵ given by $\hbar^2k^2/(2m^*)$, m^* being the *effective mass*, the radius of the Fermi sphere, k_0, is $(3\pi^2N)^{1/3}$, $S = 4\pi(3\pi^2N)^{2/3}$, $v_0 = \hbar(3\pi^2N)^{1/3}/m^*$, and therefore $\mathbf{j} = Ne^2\mathbf{E}/m^*$.

The steady increase of current does not continue indefinitely in a real metal on account of collisions between the electrons and lattice imperfections, e.g. impurity atoms or lattice waves. As a result the displacement at any region of the Fermi surface ultimately reaches a steady limit, and it is not uncommon to assume for convenience that the approach to the limit is exponential; at a time t after application of the field the displacement $\Delta\mathbf{k}$ is given by $\Delta\mathbf{k}_\infty(1-e^{-t/\tau})$, where $\Delta\mathbf{k}_\infty$ is the limiting value and τ the relaxation time, which may vary with position on the Fermi surface. If this law is obeyed, it is clear that the initial rate of increase of $\Delta\mathbf{k}$ is just $\Delta\mathbf{k}_\infty/\tau$, so that we may immediately modify (6) to give an expression for the steady current

$$\mathbf{j} = \frac{e^2}{4\pi^3\hbar^2}\int(\tau\operatorname{grad}_k \epsilon)\,\mathbf{E}.d\mathbf{S}. \qquad \ldots\ldots(8)$$

For a cubic or polycrystalline metal the conductivity takes the form, from (7),

$$\sigma \equiv \mathbf{j}/\mathbf{E} = \frac{e^2\,S\overline{v_0\tau}}{12\pi^3\hbar} = \frac{e^2\,Sl}{12\pi^3\hbar}. \qquad \ldots\ldots(9)$$

Here l, which may vary with position on the Fermi surface, is the mean free path, defined as $v_0\tau$, and $\bar{l}$ is its average value taken over the Fermi surface. We shall return later to a somewhat fuller discussion of the relaxation process.

In a magnetic field the classical interpretation of the wave-packet leads us to expect a Lorentz force $e\mathbf{H}\times\mathbf{v}$† on an electron, and this expectation is more or less (though perhaps not perfectly) justified by detailed quantal analysis. Since the velocity $\mathbf{v}$ is directed normally to a surface of constant energy, the Lorentz force always acts tangentially to such a surface, so that the energy remains

† From now on we shall work in electromagnetic units.

unchanged by the magnetic field. But the crystal momentum $\hbar\mathbf{k}$ changes at a rate equal to the Lorentz force, so that $\mathbf{k}$ travels around the orbit in reciprocal space defined by the intersection of the constant energy surface with a plane normal to $\mathbf{H}$, i.e. k_z = constant if $\mathbf{H}$ is taken parallel to the z-axis. Correspondingly, in real space the electron describes an orbit, which bears a very simple relationship to the orbit in reciprocal space. For we may write the equation of motion

$$\dot{\mathbf{k}} = (e/\hbar)\,\mathbf{H}\times\dot{\mathbf{r}}, \qquad \ldots\ldots(10)$$

from which it follows that the variations of the position vector $\mathbf{r}$ (strictly, the components of $\mathbf{r}$ in the plane z = constant) are related to those of $\mathbf{k}$ by a factor $\hbar/(eH)$, constant in constant field, and by a rotation through $\pi/2$. Thus the orbit in real space, projected on a plane normal to $\mathbf{H}$, has the same shape as the $\mathbf{k}$-orbit, but is larger by a factor $\hbar/(eH)$ and rotated through $\pi/2$. This result says nothing about the z-component of the motion, parallel to $\mathbf{H}$, which may in general be quite a complicated periodic function superposed on a mean drift; by following the variation of the z-component of $\mathbf{v}$ around the $\mathbf{k}$-orbit the nature of this motion may be elucidated in any particular case. If the energy surfaces are spherical or, more generally, axially symmetric about the z-axis, the drift velocity is constant and the electron describes a regular helix.

Since the motion in a magnetic field is periodic in the plane z = constant we may expect quantization to occur. An exact treatment for electrons in a periodic lattice has not been given, but the rigorous theory of free electrons finds a simple interpretation in the Bohr–Sommerfeld quantum theory, which can be extended heuristically (Onsager 1952) to the more general case to give results of considerable interest and value. What is involved here is an evaluation of the phase integral

$$\oint p\,dq,$$

which is set equal to a multiple of h or to $(n+\varphi)h$ where φ is an undetermined phase factor. In the phase integral the proper momentum conjugate to the position vector $\mathbf{r}$ is not $\hbar\mathbf{k}$ but $\hbar\mathbf{k}-e\mathbf{A}$, where curl $\mathbf{A}=\mathbf{H}$. If now we take $\boldsymbol{\kappa}$ and $\boldsymbol{\rho}$, measured from equivalent origins, as the components of $\mathbf{k}$ and $\mathbf{r}$ in the plane normal to $\mathbf{H}$, (10) may be integrated to take the form $\hbar\boldsymbol{\kappa} = e\mathbf{H}\times\boldsymbol{\rho}$, so that the phase integral becomes

$$\oint(\hbar\boldsymbol{\kappa}-e\mathbf{A}).\,d\boldsymbol{\rho},$$

which may be rewritten as $\oint[e\mathbf{H}.(\boldsymbol{\rho}\times d\boldsymbol{\rho})-e\mathbf{A}.\,d\boldsymbol{\rho}]$.

Since

$$\oint\boldsymbol{\rho}\times d\boldsymbol{\rho}$$

is twice the projected area of the orbit in real space and

$$\oint\mathbf{A}.\,d\boldsymbol{\rho}$$

is Φ, the flux contained within the orbit, we see that

$$\oint p\,dq = e\Phi = (n+\varphi)h.$$

This defines the permitted orbits, whose projected areas may only be $(n+\varphi)$ times $h/(eH)$; correspondingly, in reciprocal space the only permitted regions of the energy surfaces are those intersections with planes k_z = constant having areas $\mathscr{A}$ in accordance with the expression

$$\mathscr{A} = (n+\varphi)\,2\pi eH/\hbar. \qquad \ldots\ldots(11)$$

This result, which we shall call Onsager's theorem, holds the clue to the understanding of the de Haas–van Alphen effect which we shall consider in detail later. For the present, however, we note that for many purposes, particularly in the discussion of transport phenomena, it appears to be possible to distinguish features which owe their existence to quantization from those which are uninfluenced by it.

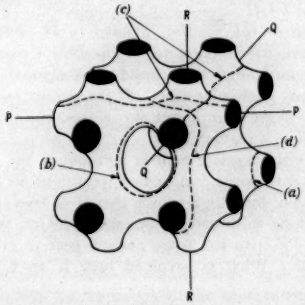


Figure 14. Multiply connected energy surface in periodically extended zone scheme showing: (*a*) electron orbit, (*b*) hole orbit, (*c*) open orbits, (*d*) extended orbit: **P**, **Q**, **R**, axes of open orbits.

The geometrical similarity of the orbits in real and reciprocal space enables us to discuss their principal features with reference to reciprocal space only, and to observe that (with one exception, which occurs when two bands are degenerate at the same value of **k**, and with which we shall not concern ourselves) all types of orbit are described in terms of sections of the periodically extended zone structure by planes normal to **H**. Thus orbits may be closed, open or extended, as illustrated in figure 14. Closed orbits may be of two types, the electron orbit (14(*a*)) which encloses states of lower energy, so that the velocity vector is directed outwards, and the hole orbit (14(*b*)) which encloses states of higher energy so that the velocity vector is directed inwards. If one looks along the direction of **H** (from South to North) an electron in an electron orbit is seen to rotate clockwise, while an electron in a hole orbit rotates anti-clockwise, as if it were positively charged. An electron in an open orbit (if orbit is the right word) describes a wavy path whose general direction is straight (14(*c*)). An extended orbit (14(*d*)) is one which extends in reciprocal space over a region which cannot be contained in one Brillouin zone, however this is placed. It should be noted, therefore, that an orbit such as 14(*b*), although it may cross the boundary of the Brillouin zone whose centre is the origin of **k**, is not an extended orbit according to our definition,

since by shifting the centre of the zone the orbit can be wholly enclosed. It is also worth noting that when the energy surface in the extended scheme is multiply-connected, as in figure 14, it is not possible to ascribe 'electron-like' or 'hole-like' character to any part of the energy surface. The same point may, with different orientations of the magnetic field, belong to electron or hole orbits, or to extended or open orbits.

Before leaving the topic of orbits let us calculate the frequency of rotation in the orbit, the *cyclotron frequency* of the particular orbit. Consider two orbits in reciprocal space, in the same section normal to **H**, being sections of the energy surfaces ϵ and $\epsilon+d\epsilon$. At any point their distance apart is $d\epsilon/|\mathrm{grad}_k\,\epsilon|$ or $d\epsilon/(\hbar v)$; so that if $d\kappa$ is an increment of length measured along the orbit, the area of the annulus between the orbits is

$$\frac{d\epsilon}{\hbar}\oint\frac{d\kappa}{v}.$$

Thus

$$\frac{d\mathscr{A}}{d\epsilon}=\frac{1}{\hbar}\oint\frac{d\kappa}{v}.$$

Now the Lorentz force on the electron is evH, so that the time taken to traverse the element $d\kappa$ is $\hbar d\kappa/evH$, and the periodic time of the orbit is

$$\frac{\hbar}{eH}\oint\frac{d\kappa}{v}\quad\text{or}\quad\frac{\hbar^2}{eH}\frac{d\mathscr{A}}{d\epsilon}.$$

We have therefore an expression for the cyclotron frequency

$$\omega_c=\frac{2\pi eH}{\hbar^2}\left(\frac{d\mathscr{A}}{d\epsilon}\right)^{-1}.$$

For free electrons having a component of wave number k in the plane normal to **H**, $\mathscr{A}=\pi k^2$ and $\epsilon=\hbar^2k^2/2m=\hbar^2\mathscr{A}/2\pi m$; so that $\omega_c=eH/m$, which is independent of the energy of the electron. In general, however, ω_c is different for each orbit. If we define the *cyclotron mass* of a given orbit by the equation $m_c^*=(\hbar^2/2\pi)(d\mathscr{A}/d\epsilon)$ we may write, analogously to the free electron case,

$$\omega_c=\frac{e}{m_c^*}H. \qquad \ldots\ldots(12)$$

It is interesting to compare (11) and (12): if the quantum number in (11) is reasonably large so that only a small fractional change of ϵ is involved in changing from one permitted level to the next, we may treat m_c^* as sensibly constant and determine the energy difference between two levels, corresponding to $\Delta\mathscr{A}$ being equal to $2\pi eH/\hbar$,

$$\Delta\epsilon=\frac{\hbar^2\omega_c}{2\pi eH}\Delta\mathscr{A}=\hbar\omega_c.$$

An oscillatory field in resonance with the orbit has thus a quantum energy which is just right for exciting a transition between permitted levels. This result, which is precisely paralleled by the Bohr orbits of hydrogen-like atoms when the quantum number is high, is a good example of the correspondence principle, and shows how, as we remarked earlier, it is often permissible to neglect quantization of orbits and to treat the problem classically.

§3. Transport Phenomena

3.1. *The Method of Trajectories*

It is necessary now to consider in a little more detail the process of attainment of equilibrium under the influence of electric and magnetic fields. A complete discussion of all the points involved would be too long for the present article, and we shall do little more than develop the particular point of view which will be most useful here. At any instant the electron distribution is described by the distribution function $f(\mathbf{r},\mathbf{k})$, which specifies the fraction of available states actually occupied by electrons of position coordinate around $\mathbf{r}$ and wave number around $\mathbf{k}$. In general, when electric fields are present $f(\mathbf{r},\mathbf{k})$ varies with position, and from a quantal point of view there is a limit to the degree of precision with which the variation of $f(\mathbf{r},\mathbf{k})$ can be specified, in accordance with the uncertainty principle. If, for example, we wish to describe a situation in which $f(\mathbf{r},\mathbf{k})$ is a rather rapid function of position, we must be able to confine the wave function of one electron to a rather small region, and this implies the construction of a wave packet from a considerable range of wave numbers. Thus we have lost the right to specify the wave number with precision, and this may lead to difficulties, particularly at very low temperatures where $f(\mathbf{r},\mathbf{k})$ changes from unity to zero over a very small range of $\mathbf{k}$ across the Fermi surface. A complete resolution of this problem is far from straightforward, but it appears to be permissible in most cases to ignore it and treat the electrons as classical particles except in so far as they obey Fermi rather than Boltzmann statistics. A not unrelated difficulty arises, as already mentioned, in a magnetic field which is strong enough that the energy separation of the permitted orbits given by (11) is comparable with kT, and again we ignore it for the present.

We proceed then on the assumption that the electrons may be pictured as particles. In the absence of collisions of the electrons with each other or with lattice defects, Liouville's theorem is applicable, the distribution function in the neighbourhood of any electron remaining constant as the electron moves under the influence of the applied fields. Since the value of f around any one electron is a function of $\mathbf{r}$, $\mathbf{k}$ and t, we have, in general,

$$\frac{df}{dt}=\left(\frac{\partial f}{\partial t}\right)_{\mathbf{r},\mathbf{k}}+\dot{\mathbf{r}}.\operatorname{grad}_{\mathbf{r}}f+\dot{\mathbf{k}}.\operatorname{grad}_{\mathbf{k}}f$$

so that in the absence of collisions, when $df/dt=0$

$$\begin{aligned}\left(\frac{\partial f}{\partial t}\right)_{\mathbf{r},\mathbf{k}}&=-\dot{\mathbf{r}}.\operatorname{grad}_{\mathbf{r}}f-\dot{\mathbf{k}}.\operatorname{grad}_{\mathbf{k}}f\\&=-\mathbf{v}.\operatorname{grad}_{\mathbf{r}}f-\frac{e}{\hbar}(\mathbf{E}+\mathbf{v}\times\mathbf{H}).\operatorname{grad}_{\mathbf{k}}f\end{aligned}$$

since $\hbar\dot{\mathbf{k}}$ is equal to the force on the electron. If the electrons are capable of colliding with impurities or with one another, there is an additional contribution to the rate of change of f, which may be written $[\partial f/\partial t]_{\text{coll}}$, so that the equation for $f(\mathbf{r},\mathbf{k},t)$ takes the form

$$\left(\frac{\partial f}{\partial t}\right)_{\mathbf{r},\mathbf{k}}+\mathbf{v}.\operatorname{grad}_{\mathbf{r}}f+\frac{e}{\hbar}(\mathbf{E}+\mathbf{v}\times\mathbf{H}).\operatorname{grad}_{\mathbf{k}}f=[\partial f/\partial t]_{\text{coll}}.$$

This is the Boltzmann transport equation in the form suitable for transport problems in metals. If it can be solved for given applied fields the electric and thermal currents immediately follow from the form of $f(\mathbf{r}, \mathbf{k}, t)$.

The Boltzmann equation is the starting point for what may be called the conventional approach to transport problems, and undoubtedly it is the best method when the collision term $[\partial f/\partial t]_{\text{coll}}$ is complicated in form. Often enough, however, a simplified form of $[\partial f/\partial t]_{\text{coll}}$ is assumed, especially when the spatial variations of $\mathbf{E}$ and $\mathbf{H}$ are sufficient in themselves to create severe mathematical difficulties, and it may then be more revealing to abandon the Boltzmann equation and go back to first principles (Shockley 1950, Chambers 1952 a). We may illustrate this by a rather special case which happens to be particularly useful for the present discussion; we consider the situation where the temperature is uniform and we are only concerned to evaluate the electric current density, on the assumption that at every point on the Fermi surface a relaxation time is definable. By this we imply that if in the absence of fields $f = f_0$, then $[\partial f/\partial t]_{\text{coll}} = -(f - f_0)/\tau$, τ being the relaxation time. Let us now follow a group of electrons through the metal, taking as our sample those electrons whose values of $\mathbf{k}$ cover a small area of the Fermi surface and have a sufficient spread of energy that the whole tail of the Fermi distribution is represented. We shall assume that the range of velocities involved is small enough for the group to spread very little, so that each electron in the group has effectively the same history. The effect of $\mathbf{H}$ is to cause the group to describe a curved trajectory, which may be calculated in principle, but not to change the energy of any electron; thus the Fermi distribution is not perturbed. Along this trajectory, however, the electrons may experience an electric field which changes their energy, and so perturbs the distribution. Consider such a perturbation $p\mathbf{E}$, caused by a field $\mathbf{E}$ at a time t_0; at a later time $t_0 + t$ collisions will have reduced this perturbation to $p\mathbf{E}\,e^{-t/\tau}$. Just the same result would have occurred if the field had had magnitude $\mathbf{E}\,e^{-t/\tau}$ and there had been no collisions. It is therefore possible to allow for the effects of collisions by a modification of the applied electric field so as to reproduce the destruction of information. Next, consider the calculation of the current density contributed by this group of electrons. Since the unperturbed distribution function f_0 carries no current, we need only consider the current due to the perturbation. Now each electron in the group suffers the same history, and therefore the effect of $\mathbf{E}$ is to shift the Fermi tail to a different energy, but otherwise leave it unaltered. Of course, groups of electrons taken from different parts of the Fermi surface will have traversed different trajectories and experienced different fields, so that the shift will vary with position on the Fermi surface. At any point the equilibrium and disturbed distributions will look as in figure 15, in which f is plotted against k_n, the component of $\mathbf{k}$ normal to the Fermi surface. If the displacement is Δ the area between the curves is independent of the shape of the Fermi tail. Thus the calculation could equally well have been carried out by assuming that the Fermi surface marked a sharp division between filled and unfilled states, as at 0°K. Since the density of states is $1/(4\pi^3)$ per unit volume of real and reciprocal spaces, the normal displacement Δ of an area dS of Fermi surface is equivalent to an increment of $\Delta\,dS/(4\pi^3)$ electrons per unit volume of real space, moving with the

Fermi velocity $\mathbf{v}_0$. Thus the increment of current density is $e\mathbf{v}_0\Delta dS/4\pi^3$ and

$$\mathbf{J} = \frac{e}{4\pi^3}\int \mathbf{v}_0 \Delta dS. \qquad \ldots\ldots(13)$$

Before considering applications of this result, one comment is perhaps useful. The method of trajectories need not be specialized to the degree we have used here; it is, for example, quite straightforward to allow τ to be a function of energy

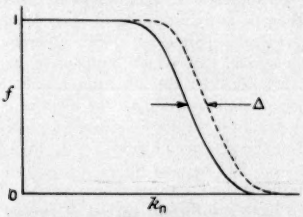


Figure 15. Displaced Fermi distribution.

as well as of position on the Fermi surface, and to consider thermal as well as electrical effects, thus deriving rather general expressions for thermal conduction and other phenomena under conditions which vary in time and space. In discussions of this *method of trajectories* (sometimes misleadingly called the *kinetic method*, as if the use of the Boltzmann equation did not involve the same kinetic principles) it is sometimes implied that the method is an approximation only suitable for rough calculations. This is true so far as concerns the assumption of a rather special form of collision term; but within the framework of this assumption the results obtained are correct integrals of the Boltzmann equation.

3.2. *Spatially Varying Electric Fields*

The use of the special form of the method of trajectories suitable for isothermal electrical conduction (which we shall call the *hard Fermi surface method*) may be illustrated by solving a problem which is relevant to much of the discussion in the following pages. Consider an infinite sample of metal within which is established a periodic electric field of frequency ω and wave number ν, having the form

$$E_i = E_{0i}\, e^{i(\omega t - \nu z)}.$$

The field will produce a current density $\mathbf{J}$ with the same periodicity, so that we may write

$$J_i = \sigma_{ij}(\omega, \nu) E_j. \qquad \ldots\ldots(14)$$

There is no difficulty in writing down a general expression for σ_{ij} in terms of the geometrical form of the Fermi surface and other relevant parameters, but we shall only consider here some rather specialized cases, particularly concentrating on the behaviour under such conditions that the mean free path is much greater than the spatial period of the field, i.e. $\nu l \gg 1$. A further simplification is effected, once the z-axis has been fixed with respect to the crystallographic axes, by choosing the x and y axes to be the principal axes in their plane. Then $\sigma_{xy} = 0$, and we need

consider only components of the types σ_{xx}, σ_{xs} and σ_{zz}. Let us take σ_{xx} first, letting **E** be directed along the x-axis, and for the moment taking ω as zero, so that $E_x = E_{0x}\,e^{-i\nu z}$. Then a typical electron at the Fermi surface passes through an alternating field, and if the free path is long, experiences very little resultant change of velocity. To calculate the change for an electron at (z_0, t_0) we note that at time t its z-coordinate was $z_0 + v_z(t-t_0)$, so that it was accelerated by an effective field

$$E_{\text{eff}}(t) = E_{0x}\exp\left\{-i(\nu z_0 + \nu v_z[t-t_0]) + \frac{t-t_0}{\tau}\right\}.$$

Hence, from the rate of change of the x-component of momentum, eE_{eff}, we derive an equation for the rate of displacement of the Fermi surface in the x-direction

$$\hbar\dot{\Delta}_x = eE_{0x}\exp\left\{-i(\nu z_0 + \nu v_z[t-t_0]) + \frac{t-t_0}{\tau}\right\},$$

which on integrating from $-\infty$ to t_0 yields the equilibrium displacement

$$\hbar\Delta_x(z_0) = \frac{e\tau}{1 - i\nu v_z\tau}E_x(z_0). \qquad \ldots\ldots(15)$$

From (15) it is clear that when $\nu v_0\tau$, i.e. $\nu l \gg 1$, the only contributions of significance to the current are from electrons moving nearly normal to z, when $v_z \ll v_0$. Thus for any section of the Fermi surface by a plane normal to the y-axis, as in figure 16, it is possible to delineate the effective points P, P′ as those

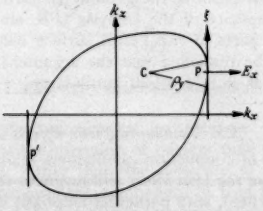


Figure 16. Effective points on section of the Fermi surface. C is the centre of curvature at P.

whose tangents are parallel to the z-axis. Only points in the near neighbourhood of the effective points have significant displacements Δ_x, and to a good approximation (which becomes exact as $\nu l \to \infty$), we may consider only this region and write $v_z = v_x\,\xi/\rho_y$, in which ξ is the coordinate k_z measured from an effective point, and ρ_y is the radius of curvature of the y-section of the Fermi surface at the effective point. The displaced area, $\int\Delta_x\,d\xi$, is now easily calculated from (15) as $\pi eE_{0x}|\rho_y|/(\hbar v_x\nu)$. If the thickness of the plane section is dk_y, the displaced area contains $eE_{0x}|\rho_y|\,dk_y/(4\pi^2\hbar v_x\nu)$ electrons for unit volume of metal, and each electron is moving in the x-direction with velocity v_x. Hence we find the current density and the conductivity component for wave number ν

$$\underset{\nu l\to\infty}{\mathscr{L}}\,\sigma_{xx}(\nu) = \frac{e^2}{4\pi^3\hbar\nu}\oint|\rho_y|\,dk_y, \qquad \ldots\ldots(16)$$

the integral being taken along the locus of effective points on the Fermi surface or surfaces. It will be seen that the expression is independent of l and v_0, and depends only on a geometrical property of the Fermi surface (Pippard 1954 a).

So long as νl is so large that the region around an effective point may be represented as having constant curvature the current is by symmetry precisely parallel to the x-axis, and $\sigma_{zx} = 0$. This result does not hold when variations of curvature are important at lower values of νl, but we may expect when νl is reasonably large that the ratio σ_{zx}/σ_{xx} will be very small. It is easy to see that as νl tends to infinity σ_{zz} also vanishes, for the only effective electrons have a negligible velocity component in the z-direction. We have thus demonstrated that in the limit of long free paths the z-axis is a principal axis of $\sigma_{ij}(\nu)$, which may be diagonalized so that only the terms σ_{xx} and σ_{yy} remain, and that these terms depend only on the geometry of the Fermi surface. These results enable the theory of the anomalous skin effect to take a very simple form, as we discuss later, and thus make the effect useful as an analytical tool.

If we do not take ω as zero the results are slightly altered in a way which is easily understood qualitatively. As an electron moves through the metal the phase of the oscillatory field changes, so that movement normal to z does not ensure a constant field. Since, however, the phase variation is that of a wave moving in the z-direction with velocity ω/ν, there will be an effective region centred about such points on the Fermi surface that $v_z = \omega/\nu$. So long, therefore, as ω is much less than $v_0\nu$, the form of the theory is hardly altered, and (16) is correct provided that ρ_y is taken at the genuine effective points. But when ω is greater than $v_0\nu$ the simplicity of the analysis fails, since there is no dominant effective region, and all parts of the Fermi surface contribute with comparable weight. It should also be remarked that the argument for the vanishing of σ_{zx} loses its validity as soon as $\nu v_0/\omega$ is significantly large.

3.3. *Galvanomagnetic Effects*

As a second example of conduction calculations relevant to our general study we consider the problem of the Hall effect and magneto-resistance, that is, uniform conduction in a magnetic field, with particular emphasis on the case when **J** and **H** are normal to one another. It is convenient to separate the effects due to electric and magnetic fields, and this can be achieved by the simple device of analysing the steady electric field into a time-sequence of impulses. We illustrate this by the free-electron model having spherical Fermi surface and constant relaxation time. Let the magnetic field be applied to an infinite sample along the z-direction, and between times $t = 0$ and $t = dt$ apply a uniform electric field E_x. This will set up a current density $dJ_x(0)$ according to the equation

$$dJ_x(0) = \frac{Ne^2}{m} E_x\, dt,$$

the whole spherical Fermi distribution being shifted in the x-direction. Now because of the field H_z the disturbed distribution precesses uniformly about the z-axis with the cyclotron frequency ω_c $(= eH/m)$, and at the same time reverts exponentially to its undisturbed form with relaxation time τ.

Thus we may write

$$\left.\begin{aligned} dJ_x(t) &= \frac{Ne^2 E_x\, dt}{m} \cos \omega_c t\, e^{-t/\tau}, \\ dJ_y(t) &= \frac{Ne^2 E_x\, dt}{m} \sin \omega_c t\, e^{-t/\tau}. \end{aligned}\right\} \quad \ldots\ldots(17)$$

If we have a steady electric field, we have a superposition of contributions of the type (17) at all epochs t, so that we obtain the steady values of J_x and J_y by integration

$$\left.\begin{aligned} J_x &= \int_{t=0}^{\infty} dJ_x(t) = \frac{Ne^2\tau}{m(1+\omega_c{}^2\tau^2)} E_x = \frac{\sigma E_x}{1+\omega_c{}^2\tau^2}, \\ J_y &= \int_{t=0}^{\infty} dJ_y(t) = \frac{Ne^2\omega_c\tau^2}{m(1+\omega_c{}^2\tau^2)} E_x = \frac{\omega_c\tau\sigma E_x}{1+\omega_c{}^2\tau^2}. \end{aligned}\right\} \quad \ldots\ldots(18)$$

The relation between **J** and **E** is shown in figure 17, in which ϕ, the *Hall angle* between **J** and **E**, is $\tan^{-1}(\omega_c\tau)$. Moreover, the triangle defined by $\sigma\mathbf{E}$ and **J** is

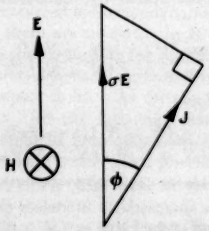


Figure 17. Illustrating Hall effect.

right-angled. In a typical experiment to measure the transverse magneto-resistance of a metal it is the direction of current flow that is determined by the geometry of the sample, rather than the direction of **E**. From figure 17 it is clear that as **H** and ϕ increase, the component of **E** in the direction of **J** remains constant at J/σ, so that there is no magneto-resistance effect in this case. The component of **E** normal to the current flow is $J/\sigma \tan\phi$ or JH/Ne, so that if we define the Hall coefficient R by the equation

$$\text{normal component of } \mathbf{E} = RHJ,$$

we see that R is simply $1/Ne$.

The vanishing of the magneto-resistance effect and the constancy of the Hall coefficient are peculiarities of the free-electron model, not observed in practice, though the nearest approximation to them occurs with sodium for which there is some evidence that the free-electron model is not a bad approximation. If we extend this analysis to more general types of Fermi surface we observe the following generalizations to be necessary :

(1) different orbits may have different cyclotron frequencies ;
(2) electron and hole orbits must be treated as having opposite signs of ω_c ;
(3) open orbits show no periodicities ;

(4) the oscillatory terms in (17) must be supplemented by harmonics of the fundamental frequency ω_c, since the disturbed distribution need not precess unchanged; in the absence of collisions it will return after each revolution to its original form;

(5) the relaxation time may vary over the Fermi surface; in general, indeed, a relaxation time will not be definable.

If all these effects are taken into account the resulting expressions are complex and unrewarding, but it is easy to demonstrate some typical properties by a simple extension of the free-electron model. We assume the Fermi surface to consist of two spherical sheets, one containing N_- electrons and the other N_+ holes, each having the same mass m and relaxation time τ, but with equal and opposite cyclotron frequencies $\pm\omega_c$. The analogues of (18) may now be written

$$J_x = \frac{\sigma E_x}{1+\omega_c^2\tau^2}, \quad J_y = \frac{N_- - N_+}{N_- + N_+}\frac{\sigma\omega_c\tau E_x}{1+\omega_c^2\tau^2},$$

from which it follows that the resistivity, determined by the component of **E** parallel to **J**, takes the form

$$\rho = \frac{1+\omega_c^2\tau^2}{\sigma}\left\{1+\left(\frac{N_- - N_+}{N_- + N_+}\right)^2\omega_c^2\tau^2\right\}^{-1}, \qquad \ldots\ldots(19)$$

and the Hall coefficient the form

$$R = \frac{(N_- - N_+)(1+\omega_c^2\tau^2)}{e\{(N_- + N_+)^2 + (N_- - N_+)^2\omega_c^2\tau^2\}}, \qquad \ldots\ldots(20)$$

in which σ is the conductivity in zero field, $(N_- + N_+)e^2\tau/m$. As the magnetic field, and consequently ω_c, is increased, ρ increases quadratically at first and then more slowly, finally levelling off at the value $\sigma^{-1}(N_- + N_+)^2/(N_- - N_+)^2$. At the same time the Hall coefficient changes from $e^{-1}(N_- - N_+)/(N_- + N_+)^2$ to $[(N_- - N_+)e]^{-1}$. It is clear that the saturation of the magneto-resistance and the Hall coefficient is possible only if the numbers of electrons and holes are different, and that if this is so the value of the field strength at which saturation begins to be complete is determined by the condition

$$\omega_c\tau \gg \left|\frac{N_- + N_+}{N_- - N_+}\right|. \qquad \ldots\ldots(21)$$

The particular results we have deduced concerning saturation and the limiting value of the Hall coefficient are of very wide validity, as we shall now demonstrate.

Let the Fermi surface be such that for the particular direction chosen for **H** (taken as the z-axis) no orbits are open. If we consider a particular section k_z = constant and apply an impulsive field $\mathbf{E}dt$ in an arbitrary direction, the displaced distribution in its subsequent precession may be expected to contribute currents of which no component need vanish. In the absence of collisions the current in any direction may be resolved into three parts

$$dJ_i = \left\{C_{ij} + \sum_n a_{ij}^{(n)}\sin n\omega_c t + \sum_n b_{ij}^{(n)}\cos n\omega_c t\right\}E_j\,dt. \qquad \ldots\ldots(22)$$

Because of collisions there will be a steady decrement of dJ_i, which may take a complicated form if the decrement in one period is large, for then differences in

relaxation at different points on the Fermi surface will show themselves. If, however, we have **H** so large that the decrement in one period is small, we may as a good approximation define a mean relaxation time $\bar{\tau}$ for the whole orbit, and hence, multiplying (22) by $e^{-t/\bar{\tau}}$ and integrating from $t = 0$ to $t = \infty$, obtain for the conductivity tensor in a steady field

$$\Delta\sigma_{ij} = \Delta J_i/E_j = C_{ij}\bar{\tau} + \sum_n \frac{a_{ij}^{(n)}\eta\omega_c\bar{\tau}^2}{1+n^2\omega_c^2\bar{\tau}^2} + \sum_n \frac{b_{ij}^{(n)}\bar{\tau}}{1+n^2\omega_c^2\bar{\tau}^2}. \qquad \ldots\ldots(23)$$

This is, of course, the contribution of one section of the Fermi surface. The complete conductivity tensor σ_{ij} involves a summation over all sections, with due account of variations in $\bar{\tau}$, ω_c and the coefficients of the various terms. For the present purpose fortunately we do not need to consider the form of (23) in detail, but only to observe that when $\omega_c\bar{\tau} \gg 1$ the three terms are respectively proportional to ω_c^0, ω_c^{-1} and ω_c^{-2}, that is to H^0, H^{-1}, H^{-2}. We now examine which of these terms may be present in each of the components of σ_{ij}, and here it is helpful to use the reciprocity relation of irreversible thermodynamics, that $\sigma_{ij}(H) = \sigma_{ji}(-H)$.† This shows immediately that there are no terms in H^{-1} in the diagonal elements of σ_{ij}, i.e. that sine terms are absent in (22) for the diagonal elements. There is no such restriction on the non-diagonal elements, in fact our analysis of the spherical model leads us to expect them. As for the constant term C_{ij}, this can only appear in the element σ_{zz}, for the following reason. As a consequence of (10) the projection of the motion of an electron on the plane $z =$ constant is a closed orbit in the absence of an electric field, provided that the orbit is closed in reciprocal space, and so no impulsive field can establish a disturbed distribution which has other than oscillatory terms in J_x or J_y. Hence the constant term vanishes in σ_{xx}, σ_{xy}, σ_{xz}, σ_{yz}, σ_{yy} and σ_{yz}, and by reciprocity in σ_{zx} and σ_{zy}. Thus if we retain only the terms of lowest order in $1/H$, as appropriate to the limiting behaviour in high fields, the form of σ_{ij} may be written

$$\sigma_{ij} = \begin{Vmatrix} A/H^2 & B/H & C/H \\ -B/H & D/H^2 & E/H \\ -C/H & -E/H & F \end{Vmatrix}. \qquad \ldots\ldots(24)$$

This is sufficient to ensure the saturation of the magneto-resistance and the Hall effect. To prove this we need to invert the conductivity tensor, since with the current direction fixed (as the x-direction) by the experimental arrangement, the measured resistivity is ρ_{xx} $(= E_x/J_x)$ and the Hall coefficient is ρ_{xy}/H $(= E_y/HJ_x)$. The result of this inversion is that

$$\rho_{xx} = \frac{\sigma_{yy}\sigma_{zz} + \sigma_{yz}^2}{\sigma_{xx}\sigma_{yy}\sigma_{zz} + \sigma_{xx}\sigma_{yz}^2 + \sigma_{yy}\sigma_{xz}^2 + \sigma_{zz}\sigma_{xy}^2} \to \frac{DF + E^2}{FB^2} \quad \text{as } H \to \infty;$$

$$\rho_{xy} = \frac{\sigma_{xy}\sigma_{zz} + \sigma_{xz}\sigma_{yz}}{\sigma_{xx}\sigma_{yy}\sigma_{zz} + \sigma_{xx}\sigma_{yz}^2 + \sigma_{yy}\sigma_{xz}^2 + \sigma_{zz}\sigma_{xy}^2} \to \frac{H}{B}, \quad \text{i.e. } \sigma_{xy}^{-1}, \quad \text{as } H \to \infty.$$

† It is not difficult by analysing an orbit in detail to demonstrate this result, so that this appeal to reciprocity need be regarded only as a means of avoiding thought.

It follows then that the resistance and Hall coefficient tend to a constant value in this general case, provided that there are no open orbits.

To calculate the limiting value of the Hall coefficient, consider a thin section of the Fermi surface bounded by parallel planes normal to **H**, and let there be s states per unit area of the section. In the presence of steady fields H_z and E_x the rate of change of energy of an electron may be written

$$\dot{\epsilon} = -ev_x E_x = \frac{E_x}{H_z}\hbar\dot{k}_y, \quad \text{since } \hbar\dot{k}_y = ev_x H_z.$$

Thus the electron does not describe in reciprocal space an orbit at constant energy ϵ_0, but one in which the energy is $\epsilon_0 + (E_x/H_z)\hbar k_y$. This second term defines the shift of the Fermi surface from equilibrium, from which the current contributed by the whole section may be written down. In particular, the y-component takes the form

$$\Delta J_y = \pm \frac{sE_x e}{H_z} \oint k_y \, dk_x,$$

in which the positive sign is to be taken if the orbit is an electron orbit, and the negative sign if it is a hole orbit. Since the integral is simply the area of the orbit, it is clear that ΔJ_y is just eE_x/H_z times the number of electrons or holes in the section, and that therefore

$$J_y = (N_- - N_+)eE_x/H_z,$$

so that the limiting value of the Hall coefficient is $[(N_- - N_+)e]^{-1}$.

This result indicates that a certain amount of information about the structure of the Fermi surface may be obtained from high-field measurements of the Hall effect, while at the same time a complete picture is unlikely to emerge. If, for example, all surfaces are closed, the numbers N_- and N_+ are independent of orientation, so that the limiting value of the Hall coefficient should be isotropic. Thus the presence of closed Fermi surfaces and something about their volumes are disclosed, but no details about their shape. It is likely that the situation in which all surfaces are closed arises most frequently with metals of even valency, in which a Brillouin zone may be nearly filled, and the next higher zone contains a few electrons; in this case $N_- = N_+$ and no saturation of the magneto-resistance is to be expected. Another, more interesting, case is that in which there is limited contact of the Fermi surface with the zone boundaries and the extended surface is multiply connected, as in figure 14. Even for those orientations which yield no open orbits the numbers of electrons and holes are likely to vary with direction of the magnetic field. The observation of a saturating magneto-resistance combined with a variable saturating Hall coefficient may prove useful in elucidating the topography of the Fermi surface, particularly in the monovalent metals which may be taken to have only one sheet to their Fermi surfaces. But the multiply-connected Fermi surface may yield highly extended and open orbits, and one must be careful not to be misled by these into misinterpretation of the results. It is worth devoting a little space to this matter, to show the influence of open orbits on the magneto-resistance and Hall coefficient.

The important thing to appreciate is that the open orbit is rather a rarity.†

† *Note added in proof.* This is not true. A fuller discussion will be found in the note at the end of this article.

An electron in an open orbit describes a path in reciprocal space which waves periodically about a straight line ; the direction of this line will be called the axis of the orbit. It is easy to delineate all possible axes as those straight lines which can pass indefinitely through the periodically extended reciprocal space without intersecting the Fermi surface. Three such lines are shown as P, Q and R in figure 14 ; it is clear that the axis is well defined and will usually coincide with a crystallographic axis of low index. This method of constructing the axis provides a necessary condition which they must satisfy, but not a sufficient condition—even if the plane normal to **H** contains one of the axes, there may still be no open orbits, as may be seen immediately from the figure. It is clear then that we can expect to find open orbits only when **H** lies in a plane normal to one of the orbit axes, and not always then. Moreover, the possibility of having simultaneously open orbits with two different axes is confined to one particular direction of **H**, normal to both ; for the surface of figure 14 this does not occur at all, since if **H** lies along the direction of P the two sets of axes of open orbits R and Q intersect and small closed orbits form in the interstices of their intersections.

We now consider the limiting behaviour in high fields when open orbits having a single axis are present. As the electron moves through reciprocal space in the general direction of the axis, it moves in real space along a similar perpendicular path. The greater the field, the less important the effects of collisions, and in the limit the direction of electron motion is strictly confined to the plane normal to the orbit axis. In this plane the response of the electrons to the electric field becomes independent of **H**. Thus the electrons describing open orbits tend to become strictly two-dimensional conductors. At the same time the components of σ_{ij} contributed by other electrons moving in closed orbits tend to zero as H^{-1} or H^{-2} (except for σ_{zz}) and the conductivity tensor has a limiting form

$$\lim_{H\to\infty} \sigma_{ij} = \begin{Vmatrix} a\sin^2\theta & a\sin\theta\cos\theta & 0 \\ a\sin\theta\cos\theta & a\cos^2\theta & 0 \\ 0 & 0 & b \end{Vmatrix},$$

in which θ is the angle between **E** and the orbit axis. The resistivity ρ_{xx} tends to infinity, as expected from the physical argument, except when $\theta = \pi/2$; more detailed analysis shows that ρ_{xx} increases as H^2. This is the same law as may be expected in the other non-saturating case, $N_- = N_+$; the two cases may possibly be distinguished experimentally through the behaviour of the Hall coefficient, but we shall not develop the argument here.

If the orientation is such that the plane normal to **H** does not quite contain any orbit axes, we may expect to find highly extended rather than completely open orbits. These will only behave differently when **H** is so large as to allow complete revolutions of the extended orbits between collisions, and under these circumstances the magneto-resistance will probably increase to a large value before finally saturating in very high fields indeed. A similar sort of effect may also occur if the collision probability is highly anisotropic. The circumstances in which this is particularly likely occur when there is rather small contact between the Fermi surface and the zone boundary, so that the connecting links between

neighbouring regions of periodically extended reciprocal space are thin. It is then possible for scattering which changes $\mathbf{k}$ by only a little to move an electron from an electron orbit to a hole orbit, and so to alter its motion completely. This possibility has not been examined in detail, but it seems likely on these grounds that an electron may have difficulty in completing an orbit which runs close to such a critical region, and we shall term such orbits *non-viable*. They may not exist, but their possible existence should be borne in mind in view of the influence they may have on delaying saturation of the magneto-resistance until exceedingly high fields have been reached. It may be pointed out that small regions of high scattering probability are difficult to detect in ordinary conductivity measurements, since the major part of the conductivity is contributed by large regions of low scattering probability. But when magnetic fields are present such small regions become much more important since they affect all electrons whose orbits pass through them.

Experimentally the situation is fairly clear, so far as the facts go, but puzzling ; for many comparatively simple metals, like copper, have a magneto-resistance which, after an initial quadratic variation, settles down to a linear variation with field which continues, apparently indefinitely, well beyond the point where saturation might have been expected (Kapitza 1929). Unfortunately, almost all the work has been done with polycrystalline samples, and it has recently been conjectured (Ziman 1958) that this is the origin of the effect, which represents an average between orientations which saturate at different values of the field as well as some which do not saturate at all because of open orbits. To justify this conjecture rigorously demands an analysis of the behaviour of a polycrystal in which each crystallite exhibits a different anisotropic conductivity, and it is probably easier to study single crystals experimentally than to carry through this analysis. In any case the information which is potentially available from such a study is far greater than anything a polycrystal can yield. It is pleasing to note that a serious experimental programme along these lines has apparently begun recently (Alekseevski, Brandt and Kostina 1958, Alekseevski and Gaidukov 1958).

It must be remembered that the whole of the foregoing analysis has been based on the semi-classical wave-packet picture, and that this is particularly suspect in high magnetic fields, where the quantization of the orbits reveals itself in the magnetic behaviour of the metal. We have seen that the condition for saturation of the magneto-resistance is that $\omega_c\tau \gg 1$; since $\Delta\epsilon$, the separation of neighbouring quantized orbits, is $\hbar\omega_c$, the condition for saturation is simply that $\Delta\epsilon \gg \hbar/\tau$, which can be seen by reference to the uncertainty principle to imply that collision broadening of the levels is small in comparison with their spacing. The condition for saturation is thus the condition for breakdown of any obvious justification for the semi-classical approach. It is probable, however, that the results of the semi-classical method have still a certain validity. A great many of the observations on magneto-resistance have been made at such temperatures that $\Delta\epsilon \ll kT$, so that through thermal excitation the characteristic oscillatory effects of quantization are not apparent in the thermodynamic properties nor in the resistance and Hall effect. When the conditions are such that $\Delta\epsilon \gtrsim kT$, the oscillatory effects (Schubnikov, de Haas–van Alphen, etc.) may appear in the magnetic and transport properties, being particularly noticeable in bismuth and graphite but detectable

in other, more typical, metals. The oscillations in resistance are, however, superposed on a general trend which appears to be of the type we have been discussing, and it seems that a clear empirical separation is possible between effects which are and are not essentially dependent on the quantization of the orbit. This conclusion may be justified theoretically, as will be indicated briefly later (see p. 214).

Many of the outstanding problems in the interpretation of the galvanomagnetic effects might well be resolved by a detailed study of high-field magnetoresistance and Hall effects in single crystals of a metal whose Fermi surface has been determined by other methods. If the observed behaviour should turn out to be intelligible by semi-classical arguments, as one would hope, the way will be open for an attempt to interpret the far more involved effects at moderate field strengths, where the detailed shape of the Fermi surface and the variation over it of velocity and relaxation time all play their parts.

§4. Analysis of Special Phenomena

4.1. *Introduction*

We have devoted some space to the galvanomagnetic effects because they are among the simplest transport phenomena which appear to be capable of yielding fairly specific information on the electronic structure of metals. Many other phenomena have from time to time been invoked as evidence supporting or refuting particular theories, and we may mention in this category the following: electronic specific heat, soft x-ray spectra, optical and infra-red absorptivity, electrical and thermal conductivities, thermo-electric effects.

The recent analysis by Heine and Cohen (1958) of the monovalent metals may be cited as an example of the use to which this type of information may be put; it may enable one to conclude that the Fermi surface of a given metal does or does not touch the zone boundary, but it hardly goes further than this. It is to be hoped that other methods, including galvanomagnetic effects, may in the course of time make unnecessary such nice weighing of indirect evidence, and we therefore devote no more space to the topics listed above, but pass straight on to the potential sources of more direct information.

The experimental methods which we shall proceed to analyse in some detail are the following:

(1) de Haas–van Alphen and related effects; variation with magnetic field of magnetic susceptibility, electrical and thermal conductivity, Hall effect, etc., which are periodic in H^{-1} and associated with the quantization of orbits in a magnetic field.

(2) Anomalous skin effect; high-frequency conductivity of a pure metal when the mean free path is much greater than the skin depth.

(3) Cyclotron resonance; the skin effect in a magnetic field, particularly under such conditions that the high-frequency field is in synchronism with the electron orbits.

(4) Magneto-acoustic effects; variations in absorption of ultrasonic waves in a magnetic field, associated with the relative sizes of the wavelength and the electron orbits, or with resonance between electron orbits and acoustic vibrations.

Of these effects the last three may be understood in quasi-classical terms, that is, classical ideas supplemented by the exclusion principle; we have, indeed, already laid the foundations of our discussion in the analysis leading up to equation (16). Only for the first group of effects must the quasi-classical approach be abandoned for the semi-classical in the sense of the type of argument used to prove Onsager's theorem. In the following discussion the effects will necessarily have to be taken one at a time, and it may be overlooked that they have one feature in common. Each has this characteristic, that the observed behaviour, under ideal conditions, depends on the properties of a small group of electrons, the particular group being selected by crystal orientation, magnetic field direction or other controllable parameters. Moreover, the behaviour depends in a comparatively simple way on

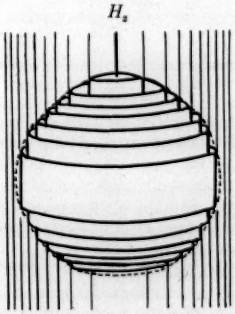


Figure 18. Permitted orbit surfaces for free electron metal. (*After* Chambers 1956 b.)

the properties of this group, so that by patient amassing and interpretation of data it becomes feasible to build up a fairly detailed picture of the variation in reciprocal space of the properties concerned. In (1), (2) and (4), for example, the property is the actual shape of the Fermi surface, $\epsilon_0(\mathbf{k})$; in (3) (and in (1) also) it is the velocity and its variation over the Fermi surface.

4.2. *The de Haas–van Alphen and Related Effects*

The de Haas–van Alphen effect has been reviewed fully by Shoenberg (1957) and Chambers (1956 b), and we shall not discuss it in great detail. Following the semi-classical approach we construct on each plane in reciprocal space normal to **H** those curves of constant energy which are permitted orbits, having areas given by equation (11), in which we shall take the phase factor φ to be $\frac{1}{2}$, being the correct value for a free-electron gas. The full diagram of permitted energy surfaces (orbit surfaces) will then consist of a set of tubes of, in general, variable cross-sectional form. Only in simple cases like the free-electron model will the tubes be cylindrical, and parallel to **H**, as in figure 18. If, for example, the energy surfaces are ellipsoidal, the orbit surfaces are elliptical cylinders, inclined in general to **H**. The degeneracy of the orbital states is such that the mean density of

states in reciprocal space is unchanged by the field. There is thus no gross change in the electronic energy with field, as is clear experimentally from the very small diamagnetism of most metals. There are, however, small changes associated with the relative dispositions of the permitted orbits and the Fermi surface. In order to show how this comes about we shall treat a not-uncommon special case, in which the surface concerned contains only a small fraction of the conduction electrons. Under these conditions the other electrons serve to maintain a constant Fermi energy, supplying or abstracting electrons as necessary so that at the absolute zero the orbit-surfaces are filled up to the points at which they coincide with the

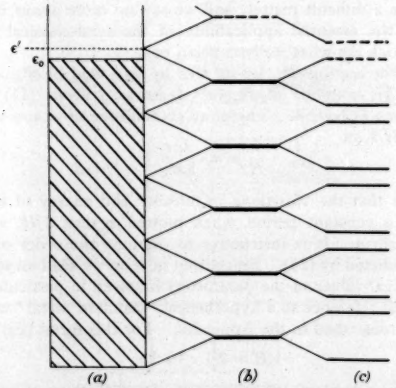


Figure 19. Energy levels in a magnetic field; (*a*) continuum in zero field, (*b*) collapse of orbital levels into highly degenerate discrete spectrum (spin neglected), (*c*) splitting of levels by spin. The levels above ϵ_0 are empty and are indicated by broken lines.

Fermi surface in zero field. If the Fermi energy is unable to stay constant the difference in behaviour is small enough that we need not discuss it here.

Consider then a section of thickness dk_z cut normal to **H**. The energy levels allowed to the electrons in this section form a fairly evenly spaced succession as in figure 19 and are just sufficiently degenerate to hold the electrons which occupy the same energy range in the absence of the field. Each of the levels is split into two by the electron spin interaction with the field, but we shall for the moment ignore this, as it is easier to correct the result for the effects of spin at the end. Since all the levels below the Fermi energy ϵ_0 are occupied, and all above are empty, the energy of the electrons in this slice amounts in all to the same as if in the absence of a field the Fermi level were at the point ϵ'; ϵ' is determined as that point, half-way between two permitted levels, which lies nearest to ϵ_0. If ϵ' and ϵ_0 coincide, the energy of the section is the same as in zero field, but otherwise it is greater. For if $\epsilon' > \epsilon_0$, the reservoir must supply electrons at energy ϵ_0 whose energy must be raised by an average amount of $\frac{1}{2}(\epsilon' - \epsilon_0)$ each, while if $\epsilon' < \epsilon_0$,

the section can hold less than its zero-field quota, the rejected electrons with energy less than ϵ_0 being raised by $\frac{1}{2}(\epsilon_0-\epsilon')$ on the average in order to enter the reservoir. Thus the effect of quantization of the orbits is to raise the mean energy of the electron assembly as well as to cause it to oscillate. The increase of mean energy is easily seen to be proportional to H^2 and since the magnetic moment M is $-d\mathscr{E}/dH$, $\mathscr{E}$ being the total energy, the average of M is negative and proportional to H, i.e. there is a steady diamagnetism, which in fact is temperature-independent. If we use this argument to calculate the *Landau diamagnetism* we arrive at the correct value for a free-electron gas, but apparently not otherwise, as pointed out by Chambers (1956 b). To take proper account of the ionic lattice is a difficult matter, and we say no more about it. There is no reason to doubt the essential applicability of the semi-classical method to the periodic terms which are what really concern us.

To return to the section dk_z, let its area in zero field be $\mathscr{A}_0$, so that for unit volume of metal it contains $\mathscr{A}_0\,dk_z/4\pi^3$ electrons. From (11) the permitted orbits have area $(n+\frac{1}{2})2\pi eH/\hbar$. Therefore coincidence of ϵ' and ϵ_0 occurs whenever $\mathscr{A}_0 = n\times 2\pi eH/\hbar$, or

$$\frac{1}{H} = n\times\frac{2\pi e}{\hbar\mathscr{A}_0}. \qquad \ldots\ldots(25)$$

This result shows that the variations in number and energy of the electrons in the section have a constant period when plotted against $1/H$, as is accurately verified by experiment. It is instructive to examine the order of magnitude of the periodicity predicted by (25). Since from now on we shall on several occasions wish to quote typical values of the parameters involved in particular experiments, we shall do so with reference to a hypothetical 'standard metal' whose definition and properties are described in the Appendix. For this metal (25) takes the form

$$1/H = 2{\cdot}1\times 10^{-9}\,n.$$

This means that in a field of 10^5 gauss, n is about 4800, so that there are 4800 permitted levels below the Fermi level. If the electrons are free, the levels are equally spaced, and the spacing is $2{\cdot}1\times 10^{-4}\epsilon_0$, which corresponds to a temperature of 13°K. This can also be seen by remembering that the separation is $\hbar\omega_c$, where ω_c is the cyclotron frequency, $1{\cdot}8\times 10^{12}$ radians/sec in the present example. If the surface is small, containing only 10^{-3} electrons per atom, $\mathscr{A}_0$ is one hundred times smaller, and n is about 48. If the effective mass of the electrons is still the free electron mass, the separation of levels is still about 13°K in a field of 10^5 gauss. Since the oscillatory effects depend for their observation on this separation being greater than the spread of electron energy due to thermal excitation, it is clear that they are likely to be observed only at rather low temperatures; this is indeed so, although when the effective cyclotron mass is small, as often happens on small Fermi surfaces, the temperature effect may not be so restrictive.

If n is fairly large, the oscillations will be reasonably evenly periodic when plotted against H, the period (i.e. the change in H which changes n by unity) being $2\pi eH^2/\hbar\mathscr{A}_0$. In figure 20 are plotted the periodic variations of the relevant properties of the section dk_z. Figure 20(a) shows ΔN, the variations in the number of electrons in the section; this varies linearly with H, with sharp breaks as an orbit crosses the Fermi level and suddenly empties into the reservoir. Since

each level contains as many states as a volume of $2\pi eHdk_z/\hbar$ would contain in zero field, the discontinuity in ΔN is $eHdk_z/2\pi^2\hbar$. At integral values of n the energy takes its minimum value, varying parabolically, as in figure 20(*b*), except for discontinuities of gradient at half-integral values of n. At such points the excess number of electrons in the section is $\pm eHdk_z/4\pi^2\hbar$, with an average excess energy of $\pm\frac{1}{2}\hbar\omega_c$ each. Hence, from (12), the range of variation of $\Delta\mathscr{E}$ is $e^2H^2dk_z/[8\pi\hbar^2(d\mathscr{A}/d\epsilon)_0]$, the derivative being taken at the Fermi surface. The

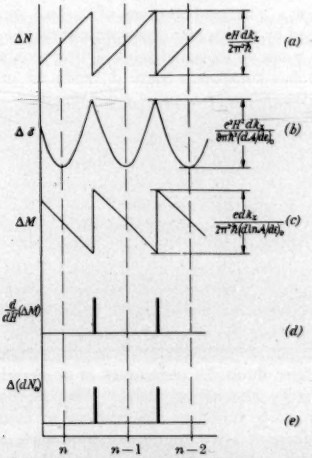


Figure 20. Illustrating origin of de Haas–van Alphen and Schubnikov effects as a consequence of the variation with H of different properties of electrons in a plane section. The vertical broken lines indicate the field strengths at which (25) is satisfied by integral values of n.

magnetic moment ΔM, being $-d\mathscr{E}/dH$, is, like ΔN, a linear function of H with discontinuities of $edk_z/[2\pi^2\hbar(d\ln\mathscr{A}/d\epsilon)_0]$, and is shown in figure 20(*c*). The other two curves do not concern us yet.

The saw tooth oscillations of ΔM owe their asperity to the sharp cut-off of the Fermi distribution at 0°K. At any other temperature thermal excitation rounds off the variations in a way which can be most easily described in terms of the Fourier components of ΔM. The amplitude of the rth harmonic must be multiplied by $X/\sinh X$, in which X is written for $2\pi^2rT/T'$, kT' being the energy separation of successive levels, i.e. $\hbar\omega_c$. For the standard metal in a field of 10^5 gauss, for which we found T' to be 13°K, the factor at 1°K is about $1{\cdot}5r/\sinh(1{\cdot}5r)$, i.e. 0·71, 0·30 and 0·10 for the fundamental, second and third harmonics respectively. A good deal of the harmonic content is thus eliminated, though a certain departure from strictly sinusoidal wave-form may still be observable, but only at rather low temperatures.

It is clear from figure 20(*c*) that the fundamental of the oscillation is of the form $\sin(\hbar\mathscr{A}_0/eH)$, so that we may take $\hbar\mathscr{A}_0/eH$ as the phase ϕ of the contribution to ΔM by the section considered. Now $\sin\phi$ is a very rapid function of $\mathscr{A}_0$, while the amplitude of the contribution, being proportional to $(d\ln\mathscr{A}/d\epsilon)_0^{-1}$, is a slow function. This enables us to sum the contributions of all sections very easily. For to a large extent they cancel one another out, except in regions where $\mathscr{A}_0$ is locally stationary with respect to k_z, i.e. where the cross section normal to **H** passes through an extremum. Such a region makes a relatively large contribution at a constant phase. The contributions of various sections at a given value of H are readily summed by means of a phase–amplitude diagram. If we measure k_z from a point where ϕ has an extremal value ϕ_0 (say a maximum), we may write $\phi = \phi_0 - ck_z^2$, where c is a constant. Since ϕ varies by many cycles in a short range of k_z, we may take $(d\ln\mathscr{A}/d\epsilon)_0$ to be substantially constant, and take as an element of the phase–amplitude diagram a vector ds of length dk_z and phase $\phi_0 - ck_z^2$. But this we know to be the construction which yields Cornu's spiral, so that we can immediately infer that the resultant of all sections is the closing vector of the spiral, and this lags in phase behind the central vector by $\pi/4$. Since, therefore, the phase of the resultant is uniquely linked to the phase of the central vector, the oscillations of ΔM for the whole surface will have the periodicity characteristic of the maximal area. If there are several extrema (maxima or minima) there will be a corresponding number of periodicities. It is not necessary that the extremal areas shall enclose filled states—extremal hole orbits should contribute equally with electron orbits to the de Haas–van Alphen effect, as should extended orbits in principle, though these may be expected to give very short periods and low amplitudes, and be very hard to observe.

We shall say little here about the techniques of observation of the effect, since they are fully discussed by Shoenberg (1957). The impulsive method which he has developed is particularly suited to observations in fields of about 10^5 gauss. The specimen, in the form of a fine monocrystalline wire, is placed in a solenoid through which is passed a heavy discharge from a bank of condensers. The field rises to about 10^5 gauss and then decays again, the whole discharge occupying about 20 milliseconds. While the field varies smoothly, the magnetic moment of the specimen oscillates very rapidly if the de Haas–van Alphen effect is present, and the oscillations may be detected by a search coil surrounding the specimen. A typical oscillogram of the signal is shown in figure 21. Two frequencies are present, but other crystals and other orientations of the same crystal may give one or three or more, depending on the number and topography of the Fermi surfaces.

The discussion of the periodicity of the de Haas–van Alphen effect has been straightforward; the amplitude is not so simple. The factors which are known to contribute to determining the amplitude in a given field are as follows:

(1) The curvature $d^2\mathscr{A}_0/dk_z^2$ at the extremum: the smaller this is, the larger the expected amplitude.

(2) The temperature: as already mentioned, this contributes to the amplitude of the fundamental frequency a factor $X/\sinh X$, where $X = 2\pi^2 kT/\hbar\omega_c$.

(3) Collision broadening of the levels: for the effect to be appreciable, the electron must have a reasonable chance of completing the orbit. If $\bar{\tau}$ is the average relaxation time around the orbit the contribution to the amplitude is a factor

$\exp(-2\pi/\omega_c\tau)$. It is, of course, important to remember that the average value of τ around an orbit may be quite different from the average over the whole Fermi surface, particularly if we are right in our supposition that near points of contact with the zone boundary there may be regions of high scattering probability. Non-viable orbits contribute no de Haas–van Alphen effect, and it would be erroneous to infer from the absence of an observable effect in any particular direction that the topography of the Fermi surface is such as to exclude closed extremal orbits.

(4) Lattice broadening of the levels : this is likely to be important only in very special orientations for which the extremal orbit has rather sharp corners as it runs close to a zone boundary. The discussions which have been given of this point

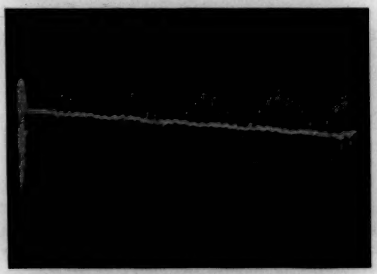


Figure 21. Oscillogram of de Haas–van Alphen effect in lead. The heavy slanting line is a trace of the variation of H. (*After* Gold 1958.)

(Harper 1955, Brailsford 1957) are physically rather obscure, but it appears that the effect may be interpreted semi-classically in the spirit of Onsager's (1952) approach. What is neglected in the latter is the fact that the form of $\epsilon(\mathbf{k})$ is determined by a specific lattice periodicity, and the electrons are not moving in a uniform background potential. The degenerate levels of the free-electron theory correspond to similar orbits around different centres in real space, and in the periodic lattice these different centres may be differently situated in the unit cell. Now the wave function of the electron is not sharply localized, so that the influence of the centring of the orbit should not be very marked. It will, however, be most marked when the orbit has sharp corners, for in the neighbourhood of these the wave packet describing the electrons becomes particularly small in real space.

(5) Electron spin : the effect of spin is to double all the levels, so that instead of a set whose energy separation is $eH\hbar/m_c^*$ we have two similar sets shifted by $geH\hbar/m$, m being the real electron mass, and g a splitting factor to allow for the possibility that spin–orbit coupling may modify the effective spin moment. The degeneracy of each set is half the degeneracy we have hitherto assumed, neglecting spin ; and since each contributes the same fundamental frequency but with a phase difference of $2\pi g m_c^*/m$, the effect is to introduce an amplitude factor $\cos(\pi g m_c^*/m)$. For the rth harmonic the factor is $\cos(r\pi g m_c^*/m)$. Very little is

known about the values of g in real metals, but it appears that in some semiconductors and in bismuth it may become very large, so that even although m_c^* may be extremely small, gm_c^*/m may be of the order of unity. This has been adduced by Cohen and Blount (1960) as an explanation for the anomalous phases of the de Haas–van Alphen oscillations in bismuth.

These are the principal theoretical contributions to the amplitude, so far as is known. There are also experimental influences which tend to reduce the amplitude, such as mosaic structure of the specimen and inhomogeneity of the field. The consequence of all this is that at present little information can be extracted from the absolute amplitude of the effect. This does not prevent the temperature variation from being used to determine m_c^*, on the assumption that the other causes of amplitude reduction are constant. The de Haas–van Alphen effect is therefore able to determine the angular variations of both the extremal cross sections and their derivatives with respect to energy. The former may be used

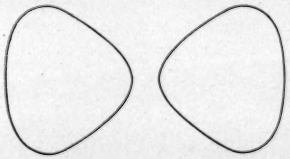


Figure 22. Fermi surface having no central extremal area.

under favourable conditions to deduce the shape of the Fermi surface uniquely, and the latter the variation of electron velocity over the Fermi surface. Lifshitz and Pogorelov (1954) have shown that if the surface is centro-symmetric and convex, so that in each orientation there is only one extremal section, passing through the centre of symmetry, there is a unique procedure for deducing the shape from the variations of periodicity, and under similar restrictions the same holds for the determination of the velocity. This is a very valuable result, but even so it does not guarantee that there will be no ambiguity of interpretation, for one cannot normally be sure that the Fermi surface satisfies the conditions of the theorem. This is especially true for a situation such as that shown in figure 22, where the Fermi surface, though centro-symmetric when considered as a whole, is split into two mirror images, neither of which has central symmetry. For any orientation of field each contributes the same periodicity, and there is no way of telling that there is not just one surface of a different shape. One could, for example, probably devise pseudo-spheres, like smoothed tetrahedra, whose extremal areas were independent of orientation, and which could occur in pairs and be indistinguishable from a single spherical surface. A second example of a surface not satisfying the conditions of the theorem of Lifshitz and Pogorelov is shown in figure 23, which is to be pictured as a dumb-bell-shaped surface of revolution. For magnetic fields in the plane of the diagram, the variation of cross section with orientation is sketched in figure 24, the branch (*a*) being the central section, and (*b*) the non-central section. At P they coalesce, since in this case there is a range of angles for which there is only a central extremum. This

situation should not be difficult to recognize in practice, especially as at the point of coalescence the second derivative $d^2\mathscr{A}_0/dk_z^2$ vanishes, and the oscillations ought to have a particularly large amplitude. Thus although the formal method of Lifshitz and Pogorelov may be inapplicable here, it should not be impossible to reconstruct the surface from the data. As a final example of ambiguity of interpretation, consider a cylindrical Fermi surface, such as might occur in a highly anisotropic metal like graphite. For a field inclined at an angle θ to the axis, the area of all sections is $\mathscr{A}_n \sec\theta$, where $\mathscr{A}_n$ is the normal cross section; rotation about the axis produces no change, and the cross-sectional shape of the cylinder is indeterminate. With these examples we leave the de Haas–van Alphen effect for the present, but shall return later to discuss some of the results which have been obtained by its use.

Closely related to the oscillations of magnetic moment are the oscillations in the transport coefficients, first observed as resistance oscillations in bismuth by

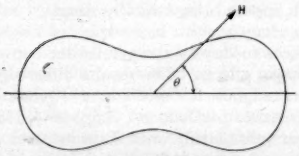


Figure 23. Dumb-bell-shaped surface having more than one extremal area.

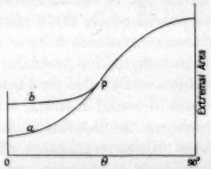


Figure 24. Variation of extremal areas with field orientation for surface of figure 23.

Schubnikov and de Haas (1930), and later observed in both transverse and longitudinal magneto-resistance, Hall effect, thermal conductivity and thermoelectric effects in a variety of substances.† Wherever comparison between any of these effects and the de Haas–van Alphen effect has been possible, the periods of the oscillations have been found to be in complete accord. We shall discuss the connection between the effects by reference in particular to the transverse magneto-resistance and Hall effect, making use of the impulsive electric field method as before. The electronic states in the magnetic field form a discrete spectrum of levels, and when the impulsive electric field Edt is applied it causes a certain number of transitions to higher states. Typically, an electron in a certain stationary state has its wave function modified so that after the impulse it has an admixture of neighbouring stationary state wave functions. In the free-electron metal only states differing in energy by one unit of $\hbar\omega_c$ will combine, and the resulting wave function describes a current-carrying state in which the current density is $Ne^2 Edt/m$,

† The following list, which is probably incomplete, indicates the variety of observations made:
Bismuth: Hall effect and magneto-resistance (Gerritsen and de Haas 1940, Alers and Webber 1953, Reynolds, Hemstreet, Leinhardt and Triantos 1954, Connell and Marcus 1957). Thermal conductivity and thermoelectric power (Steele and Babiskin 1955).
Graphite: Hall effect and magneto-resistance (Berlincourt and Steele 1955, Soule 1958 a).
Zinc: Electrical and thermal magneto-resistance (Alers 1956).
Indium arsenide: Magneto-resistance (Sladek 1958).
Note also the related oscillatory magneto-absorption of light in semiconductors (Zwerdling, Lax and Roth 1957, Lax, Roth and Zwerdling 1959).

rotating about the z-axis with frequency ω_c. In any other case the behaviour is more complex. The immediate response of the system to a sharp pulse Edt is the same as for the free-electron model, a current density Ne^2Edt/m. But this will result from the combining with the original state of many levels, some of which belong to different energy bands and give rise to very high-frequency rotational terms in the current density. If the separation of levels just above ϵ_0 is small in comparison with the band gap the contributions of inter-band transitions may be readily separated from those due to small-energy transitions, and if the former are ignored the initial current density and its subsequent oscillations are the same as given by the semi-classical calculation and expressed in the form of (22). That is to say, the discrete quantization of the levels is not in itself responsible for any significant change in the transport properties. The origin of the periodic changes in resistivity and Hall effect is to be found principally in two rather subsidiary effects of quantization, change in the number of electrons in a band and change in the relaxation time. A great deal of highly complex matter has been published on these effects, which appear to be basically simple.

The treatment which follows is an elementary analysis in the spirit of Lifshitz' (1958) approach to the problem, which seeks to display the parallelism between the Schubnikov and the de Haas–van Alphen effects. The results differ significantly from those of Lifshitz, who appears to ignore the oscillations of relaxation time which are the dominant cause of resistance oscillations. Argyres' (1958 a) calculations of the transverse effects agree substantially with Lifshitz' and are presumably open to the same criticism, although no such fault can be found with his analysis (1958 b) of longitudinal oscillatory magneto-resistance. The most recent treatment, by Adams and Holstein (1959), provides elaborate justification for the earlier results of Davydov and Pomeranchuk (1940), and for the semi-classical approach which we present here. It should be remarked that the disagreements between various workers are not concerned with the frequency of the oscillations and its relation to the Fermi surface, but only with the amplitude. The difference between Adams and Holstein on the one hand, and Lifshitz on the other, is between an oscillatory variation of resistance which is observable with difficulty and one which lies well below observability. We believe the former to represent the true state of affairs.

As with the de Haas–van Alphen effect, we consider for simplicity an assembly of electrons whose Fermi energy is maintained constant, so that the number of electrons within a section dk_z is a periodic function of field, as in figure 20(a), being related to the variations of magnetic moment by the expression

$$\Delta N = -H\left(\frac{d\ln\mathscr{A}}{d\epsilon}\right)_0 \Delta M. \qquad \ldots\ldots(26)$$

The contributions to the conductivity tensor by this section thus fluctuate as if in the absence of a magnetic field the Fermi energy were changing so as to give the variations expressed by (26). We may define an effective mobility μ_{ij} for the electrons in this section by considering the change of conductivity consequent upon a change in the number of electrons it contains, and write

$$\mu_{ij} = \partial\sigma_{ij}/\partial N.$$

This definition takes care of conductivity changes arising both directly from changes in N and indirectly from the possible variation of free path with N, but it assumes that all parts of the Fermi surface other than the section considered remain unchanged. We may now write for the variations of the contribution of the section to the conductivity from this cause

$$\Delta\sigma_{ij}^{(1)} = -H\mu_{ij}\left(\frac{d\ln\mathscr{A}}{d\epsilon}\right)_0 \Delta M. \qquad \ldots\ldots(27)$$

By the same arguments as before, most sections will contribute oscillations which cancel each other, and anything observed will arise at extremal cross sections. Thus the actually observed variations of ΔM and $\Delta\sigma_{ij}^{(1)}$ will be related by the value of the coefficient in (27) which is appropriate to an extremum. For each extremum

$$\Delta\sigma_{ij}^{(1)} = -H\Delta M(\mu_{ij}\, d\ln\mathscr{A}/d\epsilon)_{\text{ext}}. \qquad \ldots\ldots(28)$$

This is not the whole story, however, for the other parts of the Fermi surface do not remain unchanged. In particular, the density of states, taken over the whole Fermi surface, shows variations, and these are reflected in the conductivity since the probability of scattering of a given electron is proportional to the number of vacant states into which it may be scattered. A full treatment of this effect is complicated, but the general principle may be understood if we suppose that the mechanism of scattering is by impurities which scatter essentially isotropically. This means that the probability of scattering of any electron at the Fermi surface is changed in the same proportion, and this is the relative change in the total density of states. Now when the levels are quantized in a magnetic field the only permitted states just above the Fermi level are those which lie on the orbit surfaces as they cut the Fermi surface. The contribution dN_0 of a thin section dk_z to the density of states N_0 is thus zero for most values of field, since usually the area of dk_z will not be such as to allow an orbit surface. Periodically, however, the area will satisfy (11) and then dN_0 will be enormous ; in fact, dN_0 will vary with H as shown in figure 20(*e*). The area under each δ-function is readily calculated from the fact that the average density of states is only slightly variable, so that the average value of dN_0 may be taken as field-independent. Hence for each cycle,

$$\int\Delta(dN_0)\,dH = \frac{2\pi eH^2}{\hbar\mathscr{A}_0}\overline{dN_0} = \frac{eH^2}{2\pi^2\hbar}\left(\frac{d\ln\mathscr{A}}{d\epsilon}\right)_0 dk_z. \qquad \ldots\ldots(29)$$

This behaviour may be compared with that of $d(\Delta M)/dH$, which can be seen from figure 20(*d*) to take precisely the same form, the area under each δ-function being just the discontinuity in ΔM in figure 20(*c*), i.e. $e dk_z/[2\pi^2\hbar(d\ln\mathscr{A}/d\epsilon)_0]$.

Hence
$$\Delta(dN_0) = H^2\left(\frac{d\ln\mathscr{A}}{d\epsilon}\right)_0^2 \frac{d}{dH}(\Delta M). \qquad \ldots\ldots(30)$$

By the same argument as before we relate the observed variations to the extremum, and write for each oscillatory component due to this cause

$$\Delta N_0 = H^2\left(\frac{d\ln\mathscr{A}}{d\epsilon}\right)_{\text{ext}}^2 \frac{d}{dH}(\Delta M). \qquad \ldots\ldots(31)$$

Now the dependence of any component of σ_{ij} on ΔN_0 is governed by its dependence on the relaxation time. Since to observe oscillatory effects we need $\omega_c\bar{\tau}$ to be at least comparable with unity we may reasonably consider only the leading terms in the high field limit of σ_{ij} as given by (24). Reference to (23) shows that the off-diagonal terms, varying as H^{-1}, are independent of $\bar{\tau}$, while σ_{xx} and σ_{yy} vary as $\bar{\tau}^{-1}$ and σ_{zz} as $\bar{\tau}$. If then the leading term varies as H^α, it varies as $\bar{\tau}^{\alpha+1}$ and as $N_0^{-(\alpha+1)}$. The oscillatory components of σ_{ij} therefore take the form

$$\frac{\Delta\sigma_{ij}^{(2)}}{\sigma_{ij}} = -(\alpha+1)\frac{\Delta N_0}{N_0} = -(\alpha+1)\frac{H^2}{N_0}\left(\frac{d\ln\mathscr{A}}{d\epsilon}\right)^2_{\text{ext}}\frac{d}{dH}(\Delta M). \qquad \ldots\ldots(32)$$

In general, we may expect contributions of the types (28) and (32), though if there are several sheets to the Fermi surface there may be complicated interactions, since ϵ_0 will not stay constant and account must be taken of scattering from one sheet to another. If there is only one band, ϵ_0 certainly does not stay constant, but the total number of electrons does; we then expect no contribution $\Delta\sigma_{ij}^{(1)}$ (or at most a small one from the energy variation of free path) and only a contribution $\Delta\sigma_{ij}^{(2)}$. In any case, as may be seen by substituting the appropriate values for a free electron gas into (28) and (32), the ratio $\Delta\sigma_{ii}^{(2)}/\Delta\sigma_{ii}^{(1)}$ is of the order of the phase of the oscillation, which we have found to be very large indeed for Fermi surfaces holding anything approaching one electron per atom. We shall therefore concentrate on $\Delta\sigma_{ij}^{(2)}$ as giving an estimate of the magnitude of the effect which may be expected.

To estimate the magnitude of the fundamental frequency contained in ΔN_0 at the absolute zero it is convenient to work from first principles rather than to use (31). We note first that for any section dk_z the variation $\Delta(dN_0)$ is a δ-function whose fundamental has thus an amplitude equal to twice the mean value, i.e. $2\overline{dN_0}$, where $\overline{dN_0}$ is the contribution of the section to N_0 in zero magnetic field. Next we observe that, when we use the Cornu spiral to sum the contributions of all sections, the resultant has the same amplitude as if a certain central section all oscillated in phase, all other regions being neglected. The appropriate effective central section is that lying between the two planes which contribute oscillations differing in phase from the extremal oscillation by $\pi/4$. To make the matter explicit consider a spherical Fermi surface of radius k_0, so that the phase due to the extremal (central) section is $\pi\hbar k_0^2/(eH)$. The effective central section is bounded by planes $\pm k_z$ such that $\pi\hbar k_z^2/(eH) = \pi/4$, i.e. $k_z = \frac{1}{2}(eH/\hbar)^{1/2}$. The surface area of the central section is $2\pi k_0(eH/\hbar)^{1/2}$, a fraction $(eH/\hbar)^{1/2}/(2k_0)$ of the total surface area. It follows then from the foregoing argument that the relative amplitude of the fundamental oscillation is twice this, i.e.

$$\frac{\Delta N_0}{N_0} = \frac{1}{k_0}\left(\frac{eH}{\hbar}\right)^{1/2} = (2n)^{-1/2}, \qquad \ldots\ldots(33)$$

where n is the order of the oscillation as defined in (25). For the standard metal (see Appendix) n was found to be about 4800 in a field of 100 kG. We should expect the oscillations of σ_{ii} to amount to about $\pm 1\%$ if there were no factors, such as temperature or mosaic structure, reducing the effect. It is clear, therefore, that for this sort of metal we cannot expect a very striking phenomenon.

The effect is also in practice very weak in metals such as tin and zinc where the value of n is much lower. This is presumably because the oscillations are due to small portions of the Fermi surface whose total contribution to N_0 is in any case small. Where the effect really becomes marked is in semi-metals like bismuth

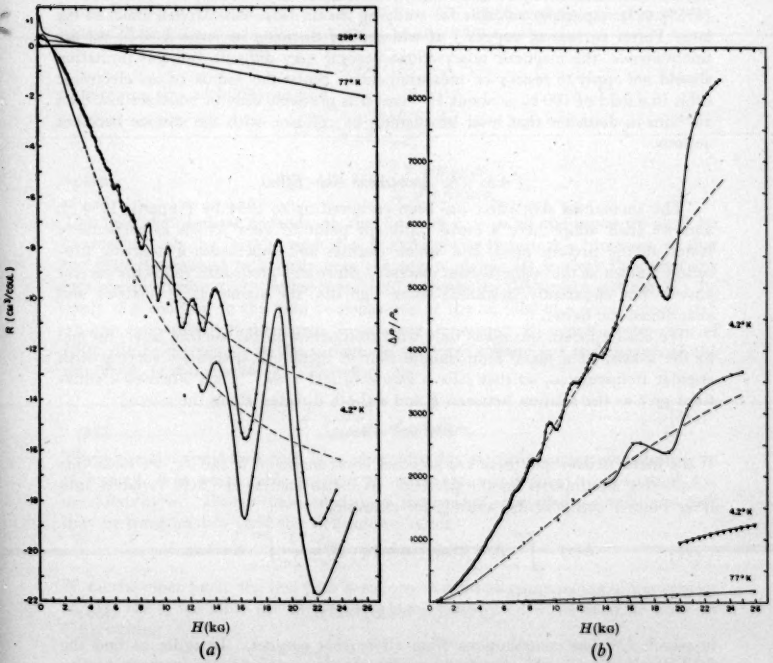


Figure 25. Field variation of (*a*) Hall coefficient and (*b*) transverse magneto-resistance in different specimens of graphite at various temperatures. The field is parallel to the *c*-axis. (*After* Soule 1958 a.)

and graphite and degenerate semiconductors like InAs and InSb where all Fermi surfaces are small. An example of the oscillations of resistivity and Hall coefficient in graphite is shown in figure 25. It will be recollected that the measured coefficients are components of the resistivity tensor ρ_{ij}, not σ_{ij}, and both diagonal and non-diagonal elements of σ_{ij} participate in the oscillatory effects.

Since the oscillations in ρ_{ij} are so intimately related to those of magnetic moment it is not to be expected that the information obtained from each should differ, and the question of which is to be preferred is solely a matter of convenience

and ease of observation. The Schubnikov effect has never been observed for any Fermi surface containing something approaching one electron per atom, perhaps only because no attempt has been made to observe it with such patience as has been lavished on the de Haas–van Alphen effect. It is possible that it may come into its own in the study of whiskers, which have already been shown by Shoenberg (1959) to be especially valuable for studying the de Haas–van Alphen effect in the large Fermi surface of copper. If whiskers of diameter as large as 0·01 cm are unobtainable, the magnetic observations become very difficult, but this limitation should not apply to resistance measurements. Since the radius of an electronic orbit in a field of 100 kG is about 10^{-4} cm, it is probably only in whiskers less than 10^{-3} cm in diameter that level broadening by collision with the surface becomes serious.

4.3. *The Anomalous Skin Effect*

The anomalous skin effect has been reviewed up to 1954 by Pippard (1954 c), and we shall adopt here a rather different point of view which is particularly suited to the present need, and which justifies and supersedes a heuristic procedure known as the *ineffectiveness concept*. Since this procedure yields the correct answer but apparently bewilders some rigorists, the sooner it is justified and abandoned the better.

We shall concern ourselves only with conduction in the surface layer, normal to the z-axis, of a plane semi-infinite slab of metal, all quantities varying with angular frequency ω, so that $\partial/\partial x = \partial/\partial y = 0$, $\partial/\partial t = i\omega$. Then Maxwell's equations give as the relation between E and J (both directed along the x-axis)

$$\partial^2 E/\partial z^2 = 4\pi i\omega J. \qquad \ldots\ldots(34)$$

If the metal is isotropic there are no other components of E and J; we make this simplifying assumption for the present. We shall analyse the field variables into their Fourier components, writing, for example,

$$E(z) = \int_{-\infty}^{\infty} E_\nu \, e^{-i\nu z}\, d\nu, \quad \text{where} \quad E_\nu = \frac{1}{2\pi}\int_{-\infty}^{\infty} E\, e^{i\nu z}\, dz,$$

so that (34) takes the form

$$\nu^2 E_\nu + 4\pi i\omega J_\nu^{\text{tot}} = 0, \qquad \ldots\ldots(35)$$

in which J_ν^{tot} has contributions from all current sources. In order to find the field distribution in the metal we require an expression for the current density produced by a given field, and this also may be written formally in terms of the Fourier components

$$J_\nu = \sigma_\nu E_\nu, \quad \text{with} \quad \sigma_\nu(-\nu) = \sigma_\nu(\nu). \qquad \ldots\ldots(36)$$

This equation defines the conductivity component σ_ν in the same way as in (14). Such a definition is immediately useful only if the metal sample is infinite in extent, and we must consider how to allow for the surface of the actual sample. A particularly simple treatment is based on one first given by Klein (1944), which is applicable when the surface is a perfect reflector of electrons. Any electron leaving the surface after a collision with it has k_x and k_y unaltered while k_z is reversed in sign. Its condition is exactly as if it had passed through a field on the

other side of the surface which was a mirror image of the field in the metal. We may therefore replace the actual problem by one in which an infinite sample of metal is subjected to a field having the property that $E(-z) = E(z)$. Such a field may be set up by a current sheet confined to the plane $z = 0$.

We therefore suppose that there is a current I per unit width on the origin plane, having a Fourier spectrum $I_\nu = I/2\pi$. The total current component consists of I_ν and the conduction current J_ν, so that (35) now reads

$$\nu^2 E_\nu + 4\pi i\omega(I/2\pi + J_\nu) = 0.$$

Combining this with (36) we find

$$E_\nu = \frac{-2i\omega I}{\nu^2 + 4\pi i\omega\sigma_\nu},$$

so that

$$E(z) = -4i\omega I \int_0^\infty \frac{\cos \nu z \, d\nu}{\nu^2 + 4\pi i\omega\sigma_\nu}.$$

The field distribution depends on the form of σ_ν; in all the cases we shall be concerned with the form is such that $E(z)$ decreases to zero as z increases, so that the field is confined to a layer near the origin plane. It is clear then that the total conduction current is just equal and opposite to the source current I. To apply this solution to the field configuration at the surface of a semi-infinite slab we can immediately write down the surface impedance Z, which is the ratio of $E(0)$, the field strength at the surface, to the total current per unit length in the skin layer, in this case $-\frac{1}{2}I$. Therefore

$$Z = 8i\omega \int_0^\infty \frac{d\nu}{\nu^2 + 4\pi i\omega\sigma_\nu}. \qquad \ldots\ldots(37)$$

When the electronic free path is so short that for any component contributing to (37) $\nu l \ll 1$, it is permissible to treat σ_ν as constant and equal to the usual d.c. conductivity σ. Under these conditions the *normal skin effect* prevails and (37) may be integrated to yield the well-known result

$$Z_n = (1+i)\sqrt{(2\pi\omega/\sigma)}. \qquad \ldots\ldots(38)$$

If, on the other hand, the free path is so long that for all components of importance in (37) $\nu l \gg 1$, we must use the result which is expressed in its extreme limit by (16), writing

$$Z_\infty = 8i\omega \int_0^\infty \frac{d\nu}{\nu^2 + is^3\nu} = \frac{8\pi\omega}{3^{\frac{1}{2}} s}(1+\sqrt{3}i), \qquad \ldots\ldots(39)$$

where

$$s^3 = \frac{\omega e^2}{\pi\hbar} \int |\rho_y| \, dk_y.$$

This result shows that in the *extreme anomalous limit*, when the limiting form (16) is applicable, the surface impedance is related to the radius of curvature of the Fermi surface around the effective zone, on which the velocity of the electron is parallel to the surface of the sample; Z_∞ is independent of the free path and of the Fermi velocity. The condition of applicability of (16) may be expressed in terms of the *resistive skin depth* δ_r, which is formally defined from Z by the relations

$$\delta \equiv \delta_r + i\delta_i \equiv Z/4\pi\omega. \qquad \ldots\ldots(40)$$

It is easy to see from Maxwell's equations that δ so defined simply measures the value of $E/(dE/dz)$ at the surface, and hence the distance within which the field would fall to zero if it varied linearly with depth. There is no implication in the definition of δ that the field falls exponentially, as it does in the normal skin effect; in fact, in the anomalous skin effect it is not exponential and is rather troublesome to evaluate. For the present, however, we need only note that the values of ν which dominate the integral (39) lie around δ_r^{-1}, so that the condition of applicability of (16) is that $l/\delta_r \gg 1$. For microwave frequencies δ_r in the anomalous limit is of the order of 10^{-5} cm for most metals, so that it is desirable to have l at least as great as 5×10^{-4} cm. This implies the use of pure metals at low temperatures, since typically l is between 10^{-6} and 10^{-5} cm at room temperature. If values of l/δ_r as great as 100 can be achieved the correction needed to obtain the limiting value of Z_∞ is only a few per cent, and can be estimated fairly reliably, as has been shown by Chambers' (1952 b) experimental confirmation of the detailed theory due to Reuter and Sondheimer (1948).

In the foregoing analysis it has been assumed implicitly that **E** and **J** are parallel to one another and to the surface of the metal. This is true for an isotropic metal, but not in general. In the extreme anomalous limit, however, the analysis holds with one simple modification. We saw (see p. 198) that in the limit the tensor $\sigma_{ij}(\nu)$ can be diagonalized, by a suitable choice of x and y axes, and that $\sigma_{zz}(\nu)$ vanishes; moreover, the principal axes depend on the Fermi surface alone and not on ν. In consequence we need not concern ourselves with possible normal components of **E** and **J**, and if we confine our attention to the principal axes in the x–y plane the result (39) is valid. For any plane surface there will be two principal values of Z_∞, which we may label $Z_{\infty x}$ and $Z_{\infty y}$. If the current flow makes an angle ϕ with the x-axis the surface impedance will follow the two-dimensional tensor law

$$Z_\infty(\phi) = Z_{\infty x}\cos^2\phi + Z_{\infty y}\sin^2\phi. \qquad \ldots\ldots(41)$$

These results were first obtained by means of the ineffectiveness concept, which is fully discussed elsewhere (Pippard 1954 a, c) and which assumes that only electrons moving nearly parallel to the surface can carry current. The present analysis justifies the physical idea behind the concept and renders it obsolete.

The assumption of specular reflection of electrons which hit the surface appears from experiment to be unsound, and is indeed theoretically improbable, for irregularities on an atomic scale are sufficient to cause large-angle diffraction of the de Broglie waves. It is almost certainly much nearer the truth to assume completely diffuse scattering at the surface, which is the only other case which has been found mathematically tractable. Even this is of a different order of difficulty from the case of specular reflection, but it has been worked out by Reuter and Sondheimer (1948), again in terms of the Fourier coefficients of the conductivity. The result takes the form

$$Z = 4\pi^2 i\omega \Big/ \int_0^\infty \ln(1 + 4\pi i\omega\sigma_r/\nu^3)\, d\nu,$$

and in the extreme anomalous limit

$$Z_\infty = \frac{\sqrt{3}\pi\omega}{s}(1 + \sqrt{3}i), \qquad \ldots\ldots(42)$$

which is 9/8 times the expression (39) for specular reflection. It is clear that the limiting value Z_∞ still depends only on the geometry of the Fermi surface and not on the mean free path or Fermi velocity.

One final point in the theory which is worth discussing is the problem of retardation effects. So long as the frequency is low and the mean free path not too long an electron can execute its path while the field is sensibly constant. In normal conduction in a uniform field the breakdown of this condition, when $\omega\tau \sim 1$, leads to a substantial change, the d.c. conductivity being replaced by $\sigma/(1+i\omega\tau)$. This effect does not occur in the extreme anomalous limit. As we have remarked in deriving the form of σ_ν, the criterion for retardation effects to be important is not that $\omega\tau$, but $\omega/v_0\nu$, shall be comparable with unity. This is equivalent in the anomalous skin effect to the statement that no serious modifications are to be expected until the electron is unable to traverse the skin depth normally in a time short compared with the period of oscillation. Since $\delta \sim 10^{-5}$ cm and $v_0 \sim 10^8$ cm/sec, for most metals the critical frequency is about 10^{13} sec^{-1}, some 10^2–10^3 times the frequencies used in practice. The effects of retardation are not negligible, but the correction to be applied amounts to no more than a few per cent. The compensation for retardation effects is managed by a shift in the position of the effective zone, different for each Fourier component of the field, so that at too high a frequency the anomalous skin effect is not governed by a rather precisely defined effective zone, and the usefulness of the phenomenon as an analytic tool is diminished. But at microwave frequencies this limitation will only be felt with semi-metals such as bismuth (Smith 1959), where the Fermi velocity is rather small and the skin depth rather large.

Up to the present the only detailed studies of the anisotropy of the surface resistance are those which have been made on tin (Fawcett 1955) and copper (Pippard 1957 a), both by calorimetric methods. The essence of the method is to let a microwave beam fall on a plane surface of metal which is mounted *in vacuo* in rather loose heat contact with the helium bath surrounding the apparatus. The resulting temperature rise measures the power absorbed and hence the surface resistance. Details of the standardization of the apparatus can be found in the original papers. We shall not discuss the results obtained at this point, but merely comment on one or two general features. First, there is no discrimination of different sheets of the Fermi surface, as in the de Haas–van Alphen effect, so that analysis of a complex metal is virtually impossible by this means alone (though it may prove useful in combination with other methods). Secondly, it is absolutely essential to use well-prepared surfaces, so that the skin layer which carries the current is in all essentials the same as the bulk of the sample. This means that all traces of mechanical strain must be eliminated by etching and electrolytic polishing, and in addition the surface must be left with no asperities on a scale comparable with the skin depth. The major experimental problem is indeed metallurgical—the microwave problems are comparatively trivial. On the other hand, there is not the need here, as in the de Haas–van Alphen effect, to obtain very perfect crystals, since small misalignments of crystallites do nothing to destroy the effect which is being observed. Finally, some comment on the analysis of the results is desirable. The measured quantity is related to an integral of the radius of curvature around the effective zone. An analytical

expression for this property in terms of the equation of the surface is very complicated and virtually useless. The process of constructing a surface which will account for the observations is thus very tedious, and in this respect the method compares rather unfavourably with the de Haas–van Alphen effect. This, indeed, is our general conclusion, but there are exceptional metals, such as copper and silver, which are susceptible to analysis by the anomalous skin effect and which show it to advantage. In a later section we shall compare the results obtained for copper by the two methods so as to display the peculiar merit of the anomalous skin effect in this case. The question of the analysis of the results has been discussed also by Chambers (1956 b) and by Kaganov and Azbel' (1955), making use of a slightly different formulation of the expressions for Z_∞. If the direction of the normal is related to that of the z-axis by the polar angles θ and ϕ, the effective zone of the Fermi surface is characterized as that for which $\theta = \pi/2$. It is not difficult to show that if ϕ is measured from the z–x plane, the two principal values of R_∞ may be written as $\sqrt{3\pi\omega/s}$ (see (42)), the principal values of s^3 being

$$\frac{\omega e^2}{\pi\hbar}\oint\frac{\cos^2\phi\,d\phi}{K(\pi/2,\phi)} \quad \text{and} \quad \frac{\omega e^2}{\pi\hbar}\oint\frac{\sin^2\phi\,d\phi}{K(\pi/2,\phi)}.$$

In these expressions $K(\theta,\phi)$ is the Gaussian curvature at (θ,ϕ), i.e. $(\rho_1\rho_2)^{-1}$, where ρ_1 and ρ_2 are the two principal radii of curvature at that point. From this it can be seen that $\overline{R_\infty^{-3}}$, the mean of $R_{\infty x}^{-3}$ and $R_{\infty y}^{-3}$, gives the mean of s^3, and hence

$$\oint\frac{d\phi}{K(\pi/2,\phi)}.$$

It is now pointed out that the theorem of Lifshitz and Pogorelov, applied to a complete set of data on the variation of this integral with crystal orientation, would enable $K(\theta,\phi)$, and hence the Fermi surface, to be uniquely determined if the Fermi surface were convex and centro-symmetric. It seems to the present author that, as hinted by Chambers, this suggested procedure is of little practical utility. The labour of obtaining complete data would be enormous: the Fermi surfaces to which the method is particularly applicable cannot be relied on to be convex: and finally, by taking the average of $R_{\infty x}^{-3}$ and $R_{\infty y}^{-3}$ a lot of the significant information may be hidden. It may indeed be more generally true that the mathematical beauty of the theorem of Lifshitz and Pogorelov encourages an overestimate of its practical value, and that a quicker way of analysing the data, whether on the anomalous skin effect or the de Haas–van Alphen effect, is by the use of the geometrical imagination and a fair amount of trial and error.

Before leaving the anomalous skin effect let us remark on its use to determine the total area of the Fermi surface. It may be shown, for example from the relation between $\overline{R_\infty^{-3}}$ and the Gaussian curvature, that the mean value of $\overline{R_\infty^{-3}}$, taken over all crystal orientations, is related to the total area S of all sheets of the Fermi surface by the expression

$$\omega^2(\overline{R_\infty^{-3}})_{\text{av}} = \frac{e^2 S}{4\cdot 3^{3/2}\pi^4\hbar}. \qquad \ldots\ldots(43)$$

If the anisotropy of R_∞ is not too great, one may approximate to $(\overline{R_\infty^{-3}})_{\text{av}}$ by $(R_\infty)_{\text{av}}^{-3}$ and thus obtain the value of S by a single measurement on a random polycrystalline sample. A comparison of the value of S with that expected from

simple models may serve as a rough check of these models. For example, if the energy discontinuities at the zone boundaries are small, the Fermi surface, when plotted in the extended reciprocal space, may approximate to a sphere, with small deformations around the regions of intersection with zone boundaries. The total area will be very similar to that of a sphere holding the same number of electrons. It is interesting, though not conclusive, that in aluminium, which Heine (1957 a) has suggested should have rather small discontinuities, the value of S is close to that of a sphere containing three electrons per atom (Chambers 1952 b, Faber and Pippard 1955). For most many-electron metals S is distinctly smaller than the free-electron value, and we may conclude that there are considerable departures from the spherical shape. The value of S may also be used in conjunction with the specific heat constant γ to estimate the mean Fermi velocity, since the density of states at the Fermi surface is $(4\pi^3\hbar)^{-1}\int dS/v_0$, i.e.

$$\gamma = \frac{k^2 S}{12\pi\hbar}\left(\frac{1}{v_0}\right). \qquad \ldots\ldots(44)$$

Such an estimate of the Fermi velocity might with advantage be compared with an estimate from the de Haas–van Alphen effect or from cyclotron resonance, for the two types of measurement involve different patterns of excitation; thermal excitations produce excited electrons and holes all over the Fermi surface, while current-carrying excitations produce excited electrons on one side and holes on the other. If interactions between excited electrons and holes play any significant part in determining the energy spectrum near the Fermi level, one may expect the velocity deduced in the two ways to differ. Thus (44) applied to detailed independent measurements of γ, S and v_0 may serve as a critical check on the assumptions of the independent-particle model.

4.4. *Cyclotron Resonance*

An understanding of cyclotron resonance in metals, which is a more complicated matter than the corresponding effect in semiconductors, involves a consideration of the effect of a magnetic field on the electrons in a metal exhibiting the anomalous skin effect. Two distinct experimental arrangements have been investigated, one in which the magnetic field is applied normally to the metal surface and the other in which it is parallel to the surface. It is the second which is of greater interest in the study of metals; it was proposed and has been analysed in detail by Azbel' and Kaner (1956, 1958). Their analysis is of very great complexity, to some extent unavoidable since an exact treatment such as they attempt involves real mathematical difficulties, altogether apart from the repellent obscurity occasioned by the generality of their model. Other published treatments of more specialized models are either avowedly (as Heine's (1957 b)) or unwittingly (as Mattis and Dresselhaus' (1958) and Rodriguez' (1958 a)) inexact. For the present we shall follow the latter's lead, since the errors involved do not seriously affect the answers, and their physical interpretation of the phenomenon is sound enough. There is, indeed, no difficulty in understanding how the resonance comes about (see figure 26). If the field is parallel to the surface the electronic orbits can return again and again to the surface layer containing the oscillatory field, provided that $\omega_c\tau \gg 1$, and if the successive traverses find the electric

field at the same phase each time, i.e. if ω is an integral multiple of ω_c, the electrons can contribute very effectively to the conduction and the impedance may be expected to take a low value. Since the cyclotron frequency, as given by (12), is proportional to H, successive resonances should be evenly spaced in H^{-1}.

We shall now assume, incorrectly, that the treatment of the anomalous skin effect given in the last section can be taken over unmodified, so that we are concerned only to evaluate the Fourier components of the conductivity, σ_ν. We shall discuss later the errors introduced by this assumption. They are not of great consequence so long as we are interested in the resonance phenomenon, since they are due to an incorrect treatment of surface scattering and the electrons involved in the resonance do not hit the surface. Even with this approximation the evaluation of $\sigma_\nu(H)$ is in general rather lengthy, but by considering only a limiting case we can shorten the work greatly. We therefore

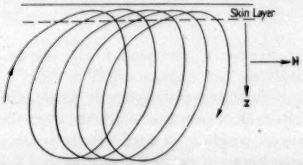


Figure 26. Electron orbit giving rise to cyclotron resonance.

choose the conditions such that the size of the electronic orbit is much larger than the wavelength of the Fourier component of the electric field, and the relaxation time is so long that $\omega_c \bar{\tau} \gg 1$. These are the conditions under which the resonance is most marked.

Take the normal to the surface of the metal as the z-axis, and let the steady magnetic field lie along the x-axis. There are two directions of microwave polarization of interest, the electric field being either parallel or normal to H. We shall consider the latter first, and discuss the behaviour of an electron in a field $E_y = E_0 e^{i(\omega t - \nu z)}$. As the electron goes round its large, not necessarily circular, orbit it suffers a rapidly alternating field from which it acquires practically no change of momentum. Only at the extremal values of z, where its velocity lies in a plane $z =$ constant, does the cancellation of contributions fail. We may therefore concentrate on the region of the orbit where z is minimal, and suppose that in this region the orbit may be represented by a circular arc of radius r; we disregard the x-component of velocity as irrelevant. The phase–amplitude diagram of the electric field acting on the electron in its trajectory forms a Cornu spiral, so that, as on p. 216, we can represent the effect of one traversal of this region by supposing the electron to move in a uniform field for such a length of arc, $\sqrt{(2\pi r/\nu)}$, as lies within one-eighth of a wavelength from the extreme point, the phase of the field being shifted by $\pi/4$ from that at the extreme. Hence if the electron passes its extreme point z at time t we may write for the change in momentum caused by one traversal

$$\hbar v_y \delta k_y = eE_0 (2\pi r/\nu)^{1/2} e^{i(\omega t - \nu z - \pi/4)}, \qquad \ldots\ldots(45)$$

where v_y is the velocity in the plane normal to H. The previous traversal, at $t-2\pi/\omega_c$, caused a similar displacement $\delta k_y'$, of which a fraction $\exp[-2\pi/(\omega_c\bar{\tau})]$ survived the circuit; the sum of all previous traversals is thus given by the expression

$$\Delta k_y = F\,\delta k_y,$$

where

$$F = \sum_{n=0}^{\infty} \exp\left[-\frac{2\pi n}{\omega_c\bar{\tau}}(1+i\omega\bar{\tau})\right] = \left\{1-\exp\left[-\frac{2\pi}{\omega_c\bar{\tau}}(1+i\omega\bar{\tau})\right]\right\}^{-1}. \quad \ldots\ldots(46)$$

It is the oscillations of the factor F which cause the observed resistance oscillations.

We must now move to reciprocal space to calculate the current at a given point z. Figure 27 shows a typical Fermi surface, of which we select for study a plane section, normal to **H**, of thickness dk_z. On this surface is drawn the effective zone, the locus of points where the normal is parallel to the surface of the metal; the effective zone cuts the section considered at Z. An electron

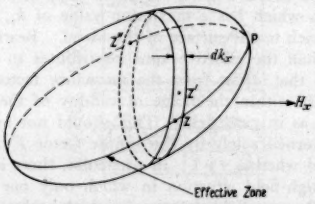


Figure 27. Illustrating calculation of cyclotron resonance.

represented by Z is moving at the extreme of its orbit, and has suffered a displacement Δk_y, as given by (46). An electron at Z' is performing a similar orbit and has suffered a displacement $\Delta k_y'$ which is the same as Δk_y in (46), except that if the electron is represented by Z' when its depth in the metal is z the extreme position of its orbit is not z, and the phase of $\Delta k_y'$ is correspondingly different from that of Δk_y. The phase–amplitude diagram of the contributions of electrons near Z to the current at z is once more a Cornu spiral, and the same rule for summation applies, the effective length of arc in reciprocal space being obtained from the effective length in real space by scaling in proportion to the radii of the orbits in the two representations. Thus the total current is the same as if all electrons in an arc of length $\rho_z(2\pi/\nu\tau)^{1/2}$ suffered the displacement Δk_y of the electron at Z, ρ_z being the radius of the Fermi surface in the plane of the section, but with a phase shift of $\pi/4$ which removes the phase shift in (45). Hence from (45) and (46), writing $1/4\pi^3$ for the density of states in reciprocal space, we have for the current density

$$dJ_y = d\sigma_{yy}(\nu,H).E_y, \quad \text{where} \quad d\sigma_{yy}(\nu,H) = F\frac{e^2|\rho_z|\,dk_z}{2\pi^2\hbar\nu}. \quad \ldots\ldots(47)$$

In deriving this result we have neglected two contributions to J_y which might appear to be important. First, there is at least one other point besides Z on the section dk_z where the electronic velocity is parallel to the surface, such as Z'' in

figure 27. But if Z represents an electron as near the surface of the metal as it can go, Z'' corresponds to the electron having gone as far into the metal as its orbit allows. If an electron represented by Z'' lies in the skin layer it will be prevented by the surface of the metal from completing its orbit, and it seems reasonable to disregard its contribution to $d\sigma_{yy}(\nu, H)$. Secondly, Z'' may correspond to the other end of an orbit of which Z represents the region near the surface; in any orbit there are two regions which contribute to J_y and we have taken only one into account. The reason for this is that the relative phase of the two contributions is very sensitive to the exact value of ν since by hypothesis $\nu r \gg 1$. In the actual problem the Fourier components of the field in the skin depth are so phased as to combine constructively near the surface and to annul one another at depths as great as r. It is therefore appropriate to regard $d\sigma_{yy}(\nu, H)$ as an average over a small range of ν, the average contribution from the far end of the orbit being zero. For similar reasons, if the Fermi surface is not wholly convex so that there is more than one effective zone, the value of ρ_x in (47) is to be taken only on that effective zone which has a maximum value of k_y, corresponding to the closest point of approach to the surface of the metal. Bearing in mind that for a convex surface only half the effective zone contributes to $d\sigma_{yy}(\nu, H)$, we see by comparison with (16) that, apart from the resonance factor F, $d\sigma_{yy}(\nu, H)$ is not only independent of H within the range of validity of the calculation, but also takes the same value as in zero field. This should not be taken to mean that $d\sigma_{yy}(\nu, H)$ will be governed solely by the simple factor F in all fields, since our treatment is only valid when $\omega_c \bar{\tau} \gg 1$; in particular, there is a difficult transition region between the high-field situation in which only one side of the effective zone contributes and the zero-field situation in which both sides contribute.

To turn now to the case where the microwave field is parallel to **H**, the calculation is readily modified to yield the result

$$d\sigma_{xx}(\nu, H) = F \frac{e^2 |\rho_y| \, dk_y}{2\pi^2 \hbar \nu}, \qquad \ldots\ldots (48)$$

where dk_y is the y-component of the length of the effective zone cut off by the section dk_z and ρ_y is the radius of the Fermi surface on the effective zone in a plane normal to the y-axis. Thus for both polarizations of the microwave field the conductivity is the same as without the magnetic field, apart from the resonance factor F.

If the effective mass and cyclotron frequency are constant over the whole Fermi surface, F will take the same value for all electrons. Since in the extreme anomalous limit the surface impedance varies as $\sigma^{-1/3}$, we may write immediately

$$Z_\infty(H) = \alpha Z_\infty(0) F^{-1/3}, \qquad \ldots\ldots (49)$$

where $Z_\infty(0)$ is given by (42). The factor α is supplied to make up for the deficiencies of the calculation which arise from taking imperfect account of the surface scattering. According to Azbel' and Kaner (1958), in the resonance region α is to a good approximation 8/9, though there are small oscillatory terms which slightly modify the amplitude of the oscillations of $Z_\infty(H)$. On account of the complex form of $Z_\infty(0)$, which has a factor $(1+\sqrt{3}i)$, the real and imaginary parts of $Z_\infty(H)$ do not exhibit symmetrical oscillations, but are distorted in the way

shown in figure 28. The distortion would hardly be noticeable experimentally for values of $\omega_c\bar{\tau}$ less than 10, but the maxima and minima in F appear somewhat displaced from the expected positions.

To examine the order of magnitude of the quantities involved, so as to find the conditions under which the effect should be observable, let us consider the standard metal (see Appendix) for which, if we use 1 cm microwaves, the fundamental resonance will occur at the quite accessible field of 11 000 gauss, the other resonances being at sub-multiples of this figure. The perimeter of the fundamental orbit is $4{\cdot}7 \times 10^{-3}$ cm, about 10^3 times larger than the skin depth, so that this condition for resonance is easily met. On the other hand, to see several oscillations of the resistance it would be desirable to have the mean free path rather larger than the perimeter of the fundamental orbit, say 10^{-2} cm. With this free path the metal would have a d.c. conductivity 2000 times better than

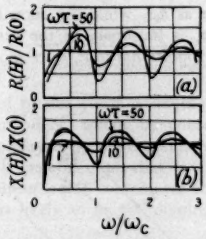


Figure 28. Theoretical variation of R and X with H in free-electron metal, for various values of $\omega_c\tau$. (*After* Azbel' and Kaner 1958.)

copper at room temperature. Such a conductivity is obtainable at the lowest temperatures with a few metals which can be prepared in a state of high purity, but the requirement is rather stringent and at present is something of a limitation to the use of this method as an analytical tool. A significant improvement would be realized by working at rather higher frequencies, since there is still a factor in hand in the ratio of orbit radius to skin depth, which varies as $\omega_c^{-\frac{1}{2}}$; but the magnetic fields required become correspondingly greater, and there are technical difficulties (probably not insuperable) in the use of microwaves of wavelength less than 4 mm, especially if resonant cavities are to be used to measure the surface impedance.

If the Fermi surface is not ellipsoidal and the effective mass is different for different orbits the interpretation of the theory becomes somewhat delicate. Azbel' and Kaner enter into the matter in great detail, but we shall do no more than discuss one or two special points of importance. Let us start with a Fermi surface something like that shown in figure 27, having a centre of symmetry. The variation of ω_c with k_z must be an even function; the point P, where the tangent plane is normal to **H** (an *elliptic limiting point* in the terminology of Azbel' and Kaner), is not especially distinguished so far as the variation of ω_c is concerned, and we may expect around P that $d\omega_c/dk_z$ will be finite and non-vanishing.

The resonance factor F will then be a smooth even function of k_x, and to determine the oscillations of the impedance it is clear from (47) and (48) that we have to evaluate integrals of the type $\int F|\rho_x|\,dk_z$ for $\sigma_{yy}(\nu,H)$ and $\int F|\rho_y|\,dk_y$ for $\sigma_{zx}(\nu,H)$. If we confine our attention to the fundamental oscillation of F (i.e. neglect the distorted form which appears at large values of $\omega_c\bar{\tau}$) the evaluation of these integrals may proceed by the use of a phase–amplitude diagram, in which an element of arc represents the amplitude (ds) and phase (ϕ) of $F|\rho_x|\,dk_x$ on an Argand diagram. The value of ϕ for any strip dk_x is seen from (46) to be $-2\pi\omega/\omega_c$, which may be written in the form

$$\left.\begin{aligned}\phi &= -\alpha/H,\\ \text{where, from (12),}\quad \alpha &= 2\pi\omega m_c^*/e,\end{aligned}\right\} \qquad \ldots\ldots(50)$$

α being a smooth even function of k_x. So far as $\sigma_{yy}(\nu,H)$ is concerned, the behaviour is dominated by the central region of the Fermi surface, around which ϕ varies no more rapidly than as k_x^2, while the contribution of a strip is sensibly proportional to dk_x, since ρ_x is stationary at the centre. The phase–amplitude diagram thus starts with ϕ proportional to s^2, once more the Cornu spiral; in this

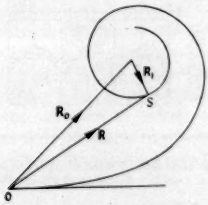


Figure 29. Cornu spiral showing contributions to cyclotron resonance.

particular application of the spiral, however, we cannot suppose that the whole length is realized, since the total difference in ϕ between the centre of the Fermi surface and the limiting point may be quite small. If it amounts to as much as 2π, the resultant oscillatory behaviour of the whole Fermi surface is easily described. Suppose that the phase–amplitude diagram is precisely a portion of the Cornu spiral, as the arc OS in figure 29, which represents $\mathbf{R}$, the resultant amplitude of $\sigma_{yy}(\nu,H)$. This resultant may be considered as dissected into two terms, $\mathbf{R}_0$ the limiting vector of the spiral, and $\mathbf{R}_1$ a correction term. Now if S lies farther than about 2π from O, $\mathbf{R}_1$ is very nearly normal to the curve at S, while, of course, $\mathbf{R}_0$ is at $\pi/4$ to the curve at O. As the magnetic field is changed, and with it all the ϕ's, S will move regularly along the spiral, getting farther from O as $\mathbf{H}$ is reduced, since from (50) $\phi = -\alpha/H$. But the resultant will always have one large term $\mathbf{R}_0$ phase-linked to the spiral at O and another $\mathbf{R}_1$ phase-linked at S. Consequently the oscillations of the conductivity will have one dominant term whose frequency is determined by the value of m_c^* at the centre of the Fermi surface, and one lesser term determined by the value of m_c^* at the limiting point. Actually this situation is rather unrealistic, since the Cornu spiral does not represent the

(s, ϕ) relationship close to the limiting point, but over-estimates the amplitude contributed by this region. The real spiral closes up more sharply, perhaps as in figure 30; the secondary contribution, from the limiting point, is then even less important, and apart from small periodic perturbations of phase and amplitude the observed oscillations will be essentially those of the central section of the Fermi surface. It is interesting that the superposition of oscillations involving

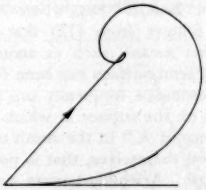


Figure 30. Modified version of figure 29.

only a fairly narrow range of phases yields a resultant in much the same way as in the de Haas–van Alphen effect, where by contrast the range is huge, the dominant role in both being played by the extremal section.

If the microwave field is polarized parallel to the magnetic field, a similar analysis shows that the limiting point contributes the region near the origin of the Cornu spiral, and this is likely to be the dominant oscillation. The central

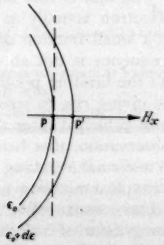


Figure 31. Cyclotron resonance at a limiting point.

section of the Fermi surface may also contribute its characteristic frequency if the magnetic field does not lie along a symmetry axis of the surface. It is not difficult to see that the section of the Fermi surface which makes no contribution is that at which the electron on the effective zone has no velocity component parallel to **H**; only if this section is also the central section will there be no oscillations from the central extremum. It should clearly be possible by measurements using different polarizations to determine which oscillations are due to limiting points and which to central sections or to other extrema of effective mass which are not located at limiting points. The possibility of singling out a particular

point on the Fermi surface is potentially very valuable, and we shall examine more closely what information it can reveal and under what conditions the effect should be observable. In figure 31 are drawn the sections of two energy surfaces, the Fermi surface ϵ_0 and a neighbouring surface $\epsilon_0+d\epsilon$, with elliptic limiting points P and P′ whose normal separation is $d\epsilon/\hbar v_0$. The tangent plane at P intersects the upper surface in an ellipse whose semi-axes are $(2\rho_{1,2}\,d\epsilon/\hbar v_0)^{1/2}$, where $\rho_{1,2}$ are the principal radii of the surface at P′. Hence the area of the ellipse is $2\pi(\rho_1\rho_2)^{1/2}\,d\epsilon/\hbar v_0$ or $2\pi d\epsilon/K^{1/2}\hbar v_0$, where K is the Gaussian curvature at the limiting point. It follows from (12) that the cyclotron frequency is $eHv_0K^{1/2}/\hbar$. If then by other means, such as anomalous skin effect studies in zero field, the shape of the Fermi surface has been found, K may be calculated; in this way the cyclotron resonance frequency can yield a value for the Fermi velocity at a particular point on the surface at which the normal is parallel to the applied field. The appearance of $K^{1/2}$ in the result emphasizes the fact that only elliptical limiting points reveal themselves, that is points at which both principal curvatures have the same sign. A wholly convex surface should show the effect at all orientations, so that in principle a complete analysis of the variation of Fermi velocity appears possible. Other cases require more care in interpretation, and we refer the interested reader to Azbel' and Kaner's thorough investigation of these points, with the mild warning that some of their suggestions for the application of cyclotron resonance, while theoretically sound enough, seem to an experimenter to be rather over-optimistic in the demands they make for purity, precision and perseverance.

As regards the conditions for observing resonance due to the limiting points, two comments may be made. The first is that one does not of course expect to observe the resonance of the electron actually at the limiting point, since it is essential to the effect that only a small fraction of the orbit shall lie in the skin depth. But if the cyclotron frequency is not an extremely rapid function of k_z there should be sections close to the limiting point having to all intents the same frequency and at the same time giving rise to sizeable orbits, at any rate for the larger Fermi surfaces which are of principal interest. The very size of the orbits, however, leads to the second observation, that between two traverses of the skin layer the electron travels quite a distance (taking $1{\cdot}4\times10^{8}$ cm sec^{-1} as a typical velocity and 3×10^{-11} second as a typical microwave period we see that this distance is $4{\cdot}2\times10^{-3}$ cm at the fundamental resonance). When a limiting point is being studied the electrons have a component of velocity parallel to $\mathbf{H}$ which is very nearly the Fermi velocity; thus the paths of the electrons are open helices having the direction of $\mathbf{H}$ as axis and a pitch of about 4×10^{-3} cm. If $\mathbf{H}$ is not precisely parallel to the surface of the metal the depths at which successive traverses take place will not be the same, and this will prevent the observation of cyclotron resonance. Since typically the pitch of the helix is 500 times the skin depth, alignment to a few minutes of arc or even better is desirable if the limiting point resonance is to be observed. It should be remembered that the surface impedance is complex, which implies that the phase of the oscillation varies with depth. Even if successive traverses lie well within the skin depth there may still be a displacement of the resonance by this phase shift. The limiting points occur in pairs at opposite ends of the Fermi surface, and have opposite displacements;

lack of exact parallelism may therefore show itself by the appearance of two resonances of equal amplitude beating with one another, or, when this effect is averaged over all contributing orbits, a diminution in the amplitude of the oscillations. The extreme sensitivity of the limiting point resonance to alignment makes it doubtful if it has ever been observed. It is more likely that the recorded resonances are due either to ellipsoidal Fermi surfaces of constant mass, which resonate all of a piece, such as those in bismuth which Aubrey and Chambers (1957) have studied in some detail (figure 32), or to sections of extremal mass in

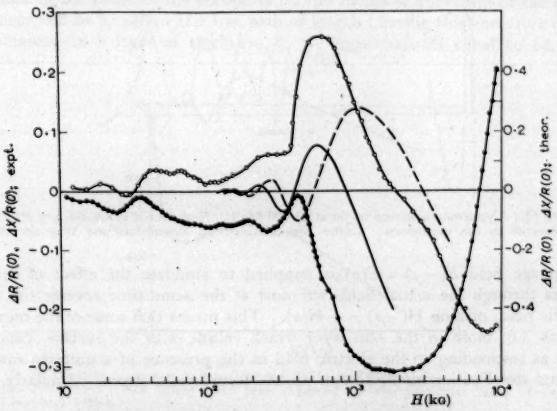


Figure 32. Cyclotron resonance in bismuth at 9400 Mc/s;
———, ● $\Delta R/R(0)$, — — —, ○ $\Delta X/R(0)$.
The theoretical curves are calculated for $m_e^* = 0{\cdot}11$ electron masses and $\omega\tau = 3$. (*After* Aubrey and Chambers 1957.)

large surfaces, as we venture to interpret the observations on tin (figure 33). This does not mean that the limiting points are of no value, but only that a far more careful investigation is needed before their worth can be assessed. The degree of parallelism required is certainly formidable, especially as strain-free electropolished surfaces are not normally very flat. Central sections, on the other hand, do not demand such a high degree of parallelism, since the mean velocity parallel to **H** is zero. The variation of the resonance with field orientation should prove of value in diagnosing the origin of the resonance. It is, however, rather unprofitable to speculate in detail along these lines until more systematic experimental evidence is available.

The foregoing discussion has been based on the assumption that the whole problem may be treated in terms of the Fourier components of the field and the response of the electrons to them, without explicit reference to the boundary conditions. This proved to be possible in zero magnetic field since in the two cases chosen, specular reflection and diffuse scattering at the surface, the electrons

leaving the surface could be treated as if they had come from the exterior of the metal, having passed through a suitable 'image field'; in the former case the image field was defined by the equation $E(-z) = E(z)$, in the latter $E(-z) = 0$. But when there is a magnetic field present this particular device becomes unworkable. The trajectory of an electron reflected specularly is as shown in figure 34;

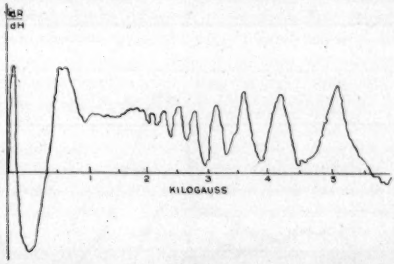


Figure 33. Cyclotron resonance in tin at 24 000 Mc/s. Note that it is dR/dH, not R, which is measured in this experiment. (*After* Kip, Langenberg, Rosenblum and Wagoner 1957.)

if an image field $E(-z) = E(z)$ is supplied to simulate the effect of multiple passages through the actual field, we must at the same time reverse the steady magnetic field, making $\mathbf{H}(-z) = -\mathbf{H}(z)$. This means that some of the important electrons, i.e. those in the skin layer which collide with the surface, cannot be treated as responding to the electric field in the presence of a uniform magnetic field, and the Fourier method in its simple form breaks down. Similarly, if we

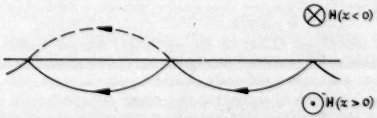


Figure 34. Electron trajectory for specular surface reflection, showing equivalent trajectory in an infinite sample.

replace diffuse boundary scattering by the assumption that $E(-z) = 0$ in an infinite medium, unless at the same time we make $\mathbf{H}(-z)$ vanish, we allow the possibility of multiple traverses between successive collisions of the skin layer by electrons which in fact collide with the surface. It appears, then, that the simple extension of the anomalous skin effect theory which we have adopted is not justified, but fortunately in the resonance region the only errors incurred are of the numerical constants, not the qualitative features, and are in any case quite small. There is, however, a phenomenon which needs the full treatment to find a satisfactory explanation, and this is a drop in surface resistance in quite small fields, as illustrated by figure 35. In this curve the hump is believed to be a rather ill-marked resonance, but the initial fall (which should surely start as H^2 in small enough

fields) is not associated with the resonance. The same effect has been seen in tin at a frequency as low as 3000 Mc/s (Pippard, unpublished), certainly too low for a resonance to be observed in the particular sample used. According to Azbel' and Kaner, whose theory takes explicit account of the surface scattering difficulty, the characteristic field strengths for this effect are those which bend the electron trajectories sufficiently to prevent any free path being completed wholly within the skin layer. Referring back to figure 27, we see that a typical electron moving parallel to the surface is represented by Z, and if the radius of curvature in the plane of the section is ρ_z, the radius of curvature of the orbit in real space will be $\hbar\rho_z/eH$. If a free path of length l having this curvature can just be contained in a layer of thickness δ_s, we must have H equal to $8\delta_s\hbar\rho_z/el^2$.

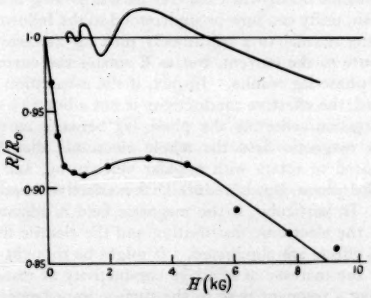


Figure 35. Cyclotron resonance in tin (Fawcett 1956) at 24 000 Mc/s. The upper curve shows the oscillations which might have been expected for the particular value of $\omega\tau$. (*After* Chambers 1956 b.)

Typically at microwave frequencies, this might be about $3 \times 10^{-5}/l^2$ gauss, so that for rather pure copper at the lowest temperatures, where $l \sim 10^{-3}$ cm, a field of 30 gauss is enough to affect the resistance appreciably. For the purest tin available the value of l is nearly 10^{-1} cm, and a field of only 3 milligauss should be sufficient. But this corresponds to an orbit radius of about 250 cm, very much more than local radii of curvature on normal electropolished surfaces; we may perhaps doubt whether the theory is valid under these circumstances. If, however, the theory may be safely applied when the orbit radius is reasonably small, there seems to be a possibility of making a direct measurement of the mean free path on the effective zone by a systematic application of this effect. Considerable evaluation of the Azbel'–Kaner theory is needed in order to discover the predicted variation of R with H (they give only the initial quadratic variation for a metal with a single ellipsoidal Fermi surface), for it is certain that a careful quantitative comparison between theory and experiment is needed before any results obtained can be regarded as reliable. It would also be of interest to attempt a solution of the theoretical problem with different boundary conditions (e.g. specular reflection), for it is likely that the behaviour is sensitive to the nature of the surface scattering, and it might be possible to use the effect to demonstrate, even more

clearly than does Chambers' (1952 b) careful work, that the surface scattering is in effect completely diffuse.

We must now turn briefly to another phenomenon, also called cyclotron resonance (Galt *et al.* 1955, 1956, 1959), which under rather special conditions can yield information on the effective masses. In this the experimental arrangement is such that the magnetic field is applied normally to the surface of the metal. Let us see what effect such a field can have on the surface impedance of a free-electron metal. It is convenient, both in order to understand the effect and to get the most from it experimentally, to have the microwave field circularly polarized, so that the phase angle of any field component represents the actual orientation of the field, and all field variables rotate about the z-axis. The retardation effects which occur when the free path is so long that $\omega\tau$ is comparable with, or greater than, unity can now be understood in the following way : electrons moving in the plane normal to z continually pick up momentum from the field and hence contribute to the current, but as $\mathbf{E}$ rotates the current already created does not and so a phase lag results. In fact, if the z-variation of $\mathbf{E}$ is sufficiently slow to be neglected, the effective conductivity is not σ but $\sigma/(1+i\omega\tau)$, the complex nature of the expression reflecting the phase lag between current and field. In the presence of a magnetic field the whole electronic distribution, and hence the current, is caused to rotate with angular velocity ω_c, and if ω and ω_c have the same sense the phase lag is reduced, the effective conductivity becoming $\sigma/[1+i(\omega-\omega_c)\tau]$. In particular, if the magnetic field is adjusted to the resonant condition, $\omega=\omega_c$, the electronic distribution and the electric field rotate in phase and all retardation effects are eliminated. It might be thought that if $\omega\tau$ is much greater than unity the increase of effective conductivity at resonance would show up as something like a resonant peak in the surface impedance, but this is not so. To continue with the present example, if the z-variation of $\mathbf{E}$ is slow we may use the normal expression (38) for the surface impedance, and write

$$Z(H)=R+iX=\left\{\frac{4\pi i\omega}{\sigma}[1+i(\omega-\omega_c)\tau]\right\}^{\frac{1}{2}}. \qquad \ldots\ldots(51)$$

The real and imaginary parts of Z are monotonic functions of ω_c, as shown in figure 36 ; only $|Z|$ shows anything like resonant behaviour, and this unfortunately is not easily measured directly. The resistance, which is easiest to measure, shows an inflection at the point of resonance. One would hesitate to lean heavily on the determination of a point of inflection for deriving an important parameter, since such a point may readily be shifted by quite minor secondary effects. If there is more than one Fermi surface, and the normal skin effect theory is still applicable, the behaviour of $R(H)$ may be more complicated and peaks may even appear, as shown by the experimental results for bismuth (figure 37). The positions of the various features of such curves as this may be interpreted in terms of Shoenberg's (1957) ellipsoidal model of the bismuth Fermi surface, but we may be permitted to doubt if this interpretation would have been arrived at without the aid of Shoenberg's very specific model. It is, indeed, a serious criticism of this experiment as an analytical tool that the recognizable features of the curves appear at positions which are determined by an implicit relation between all the different contributing electrons, and that this relation is bad enough for ellipsoidal

Fermi surfaces which represent the most complicated case yet analysed theoretically. It is nevertheless possible, as recent measurements on zinc (Galt, Merritt, Yager and Dail 1959) have shown, to use this experimental arrangement in conjunction with Azbel'–Kaner resonance with advantage. The latter determines the effective mass, and hence the field strength at which resonance should occur in a normal field. If circular polarization is now used, it proves to be possible to see sometimes whether the resonance occurs with positive or negative field, and hence to ascribe the resonance to electron or hole orbits. This is something which is not

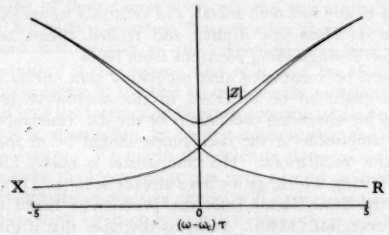


Figure 36. Variation of R, X and $|Z|$ from equation (51).

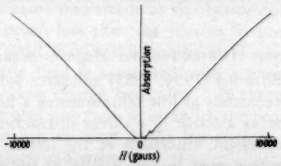


Figure 37. Field variation of absorption due to bismuth in a circularly polarized resonant cavity at 24 000 Mc/s. (*After* Galt, Yager, Merritt, Cetlin and Brailsford 1959.)

revealed directly by any other method, except possibly the Kjeldaas ultrasonic experiment which we discuss later (p. 252), and it would establish normal field resonance as a tool of great utility if there were the possibility of general application.

The method is, however, subject to the serious criticism that the phenomenon may usually be unobservable (Chambers 1956 c). We have seen how at the resonant condition retardation effects are neutralized; the magnitude of the observed effect is thus dependent on the extent to which the impedance in zero field is modified by retardation effects. But when the anomalous skin effect is well-developed, retardation effects are of little importance (see p. 221) in most metals at frequencies less than 10^{12} sec^{-1}. At microwave frequencies, as the field is changed all that happens is that the position of the effective zone shifts slightly so as to bring into prominence those electrons which have the appropriate z-component of velocity to move in phase with the rotating field. We expect to see very little, then, with most metals except at very high, long infra-red frequencies (and in correspondingly high fields of several hundred kilogauss). The

most likely exceptions to this negative conclusion are semi-metals such as bismuth and graphite which even at microwave frequencies show pronounced retardation effects, metals such as zinc in which there are small as well as large sheets to the Fermi surface, the former playing a significant role in conduction, and metals with highly distorted Fermi surfaces for which the complicated orbits in a magnetic field may lead to excessive boundary scattering of the most effective electrons and hence to a modification of the surface impedance. But the theory of the last, which is likely to be rather difficult, has not been worked out. We are therefore inclined to see very little prospect of this method making a significant contribution to the analysis of any but semi-metals and relatively minor portions of metals, and for these the de Haas–van Alphen and related effects seem vastly more powerful, except for distinguishing electrons from holes.

Finally, it should be mentioned that oscillatory skin effects analogous to the Schubnikov effect ought to be observed in the microwave impedance, under conditions suitable for observing oscillations of the d.c. resistance. According to Azbel' (1958) the amplitude of the oscillations should be of the order of ϵ_0/kT greater than the d.c. oscillations. His comparison is made, however, with the theory of Lifshitz (1958), which, as we have already seen (p. 214), predicts oscillations smaller by the same factor than the predictions based on the views of Davydov and Pomeranchuk (1940). It seems therefore that if the latter is correct there is little point in looking for the skin effect oscillations when a d.c. measurement should do as well. We shall therefore not discuss this suggested experiment further.

4.5. *Ultrasonic Attenuation and Magneto-acoustic Effects*

The experiments of Bömmel (1954, 1955) and later workers (see Mason 1958) have revealed that in pure metals at low temperatures a new mechanism of ultrasonic attenuation appears, as a result of a direct interaction between the conduction electrons and the ultrasonic vibrations of the lattice, and that in propitious circumstances the attenuation may be an oscillatory function of the strength of an applied magnetic field. The analysis of these effects leads us to what are potentially the most complicated of all the phenomena to be considered in this review. The general case of the interactions between the lattice vibrations of an anisotropic solid and the electrons lying on an arbitrary Fermi surface, in the presence of a magnetic field, conjures up a vision of almost unlimited complexity. The principal effects can, however, be made clear with the aid of the free-electron model of an elastically isotropic solid, which is the only model so far treated in print. What little has been published on the theory is at times somewhat misleading (Pippard 1955, 1957 b, Steinberg 1958 a, b, c, Rodriguez 1958 b, Kjeldaas 1959, Kjeldaas and Holstein 1959, see also Mason 1958 for other references), while as for the experimental results, they are so suggestive, and yet so tantalizingly scrappy, as to do little more than point to the need for systematic study, with the virtual certainty that something of interest will emerge. It is therefore worth while devoting some space to a detailed discussion of the free-electron model. But before doing this we shall take the opportunity of suggesting a standard notation which will serve to describe any particular experimental arrangement. So long as the metal is isotropic, the disposition of the experiment is defined by

the polarization of the wave and the direction of the applied magnetic field; if we take a Cartesian system of axes, with unit vectors $\mathbf{i}, \mathbf{j}, \mathbf{k}$, the experiment is defined by two vectors $(\mathbf{X}, \mathbf{Y})$ by which it is to be understood that the wave vector $\mathbf{q}$ is directed parallel to $\mathbf{i}$, the particle velocity $\mathbf{u}$ in the wave parallel to $\mathbf{X}$, and the magnetic field $\mathbf{H}$ parallel to $\mathbf{Y}$. Thus $(\mathbf{j}, \mathbf{k})$ represents a transverse wave in a transverse magnetic field which is normal both to the propagation direction and to the particle velocity; we shall be concerned primarily with six special cases of interest $(\mathbf{i}, 0)$, $(\mathbf{j}, 0)$, $(\mathbf{i}, \mathbf{j})$, $(\mathbf{j}, \mathbf{j})$, $(\mathbf{j}, \mathbf{k})$ and $(\mathbf{j}, \mathbf{i})$. To deal with real anisotropic metals (and of course even cubic metals may be highly anisotropic elastically as well as in the shape of their Fermi surfaces), the above notation is insufficient. We propose that for such a situation three vectors should be used, referred to specified axes fixed with respect to the crystallographic axis; the first vector defining the direction of $\mathbf{q}$, the second that of $\mathbf{u}$, and the third that of $\mathbf{H}$. We have at present no occasion to use this more elaborate notation.

To understand how the interaction between acoustic vibrations and the conduction electrons leads to attenuation, let us first suppose that the electrons suffer no collisions with the lattice. We might then picture the wave as affecting only the positive lattice, leaving the electrons at rest with a uniform density. If the wave is longitudinal the density of positive charge will vary periodically, so that in the absence of a compensating variation of negative charge there will be set up powerful longitudinal electric fields, which will force the electrons into motion. In fact, of course, only the minutest degree of charge imbalance is needed to keep the electronic current at the same magnitude as the lattice current (provided that the ultrasonic frequency is much less than the plasma frequency $\omega_p = c(4\pi Ne^2/m)^{1/2}$; since $\omega_p \sim 10^{16}$ sec^{-1} in most metals this condition is easily met). We conclude that a longitudinal electric field is associated with a longitudinal wave, having a magnitude sufficient to keep charge neutrality. If there are collisions between electrons and lattice they also may help to establish the required electronic current, but whatever the mechanism we may divide the current into contributions from the lattice and the electrons, and write

$$\mathbf{J}_{\text{latt}} + \mathbf{J}_{\text{el}} = 0. \qquad \ldots\ldots(52)$$

If the wave is transverse there is no tendency for space charges to be established, but now the movement of lattice without electrons causes a transverse current and hence a magnetic field associated with the wave. By induction an electric field is produced and it is this which pulls the electrons into motion with the lattice. The need for 'current neutrality' is not nearly so strong as for charge neutrality, the condition being that the frequency shall be much less than that at which the skin depth for alternating fields becomes equal to the wavelength. At lower frequencies than this the skin depth is smaller than the wavelength and each part of the wave is, as it were, screened from the other parts and transmits no magnetic effects; this can only be achieved if there is no net current. For most metals the critical frequency is around 10^3 Mc/s, so that the ultrasonic frequencies in use at present, up to about 100 Mc/s, are reasonably well within the condition for current neutrality. We shall assume then that (52) holds also for transverse waves, but it should be remembered that a more thorough treatment may be required in special cases (such as bismuth, where the wave velocity is low

and the skin depth high). It is worth remarking that if the frequency is well above the critical value, as with the thermally excited phonons at all but the very lowest temperatures ($< 0{\cdot}1°\text{K}$), the electrons are not drawn into motion with the lattice, and lose their power of attenuating transverse waves.

The electric field which effects current neutrality is the link between lattice and electrons which is responsible for transferring energy from the former to the latter and so attenuating the wave. If the peak value of the electric field is $\mathbf{E}$, the mean rate of increase of energy of the electron assembly, per unit volume, is given by

$$\dot{W} = \tfrac{1}{2}\mathscr{R}(\mathbf{E}^*.\mathbf{J}_{\text{el}}), \qquad \ldots\ldots(53)$$

and because of (52) this is the mean rate of decrease of energy of the lattice. Thus (53) represents the rate at which mechanical energy is converted into excitation energy of the electron assembly. Inelastic collisions transfer this energy back to the lattice as heat, but this part of the process does not concern us—the rate at which it occurs does not affect the attenuation. We must not ignore the possibility that collisions between the lattice and the electrons† may also dissipate the mechanical energy of the wave, though in fact it is easily seen that this is not so provided (52) holds. The reason is that at any point the electronic current, as seen from a frame of reference moving with the lattice, vanishes and with it the total momentum of the electron assembly. The statistical effect of collisions between electrons and scattering centres moving with the lattice is to bring the electron assembly into equilibrium with the moving lattice. Since the equilibrium state is one of zero momentum, there is no net momentum transfer in the collision processes, and hence no force of reaction between the electrons and the lattice, other than that due to the electric field. In calculating the attenuation then, we need only consider the influence of the electric field, as expressed in (53).

If the peak value of the particle velocity in the wave is $\mathbf{u}$, the lattice current density $\mathbf{J}_{\text{el}}$ is $-Ne\mathbf{u}$, where N is the number of free electrons per unit volume, and the energy per unit volume in the wave, W, is $\frac{1}{2}Mu^2$, where M is the density (we shall not concern ourselves with the minute difference between the lattice density and the density of the metal). Thus

$$\frac{\dot{W}}{W} = \frac{Ne}{Mu^2}\mathscr{R}(\mathbf{E}^*.\mathbf{u}), \qquad \ldots\ldots(54)$$

from which it is seen that the characteristic time for loss of mechanical energy is the reciprocal of the right-hand side of (54), and the extinction distance is this multiplied by the wave velocity v_s; alternatively, if the wave takes the form

$$\mathbf{u} \equiv \mathbf{u}\, e^{-i\alpha z}\, e^{i(\omega t - qz)}, \qquad \ldots\ldots(55)$$

the attenuation constant α is the reciprocal of the extinction distance, or

$$\alpha = \frac{Ne}{Mu^2 v_s}\mathscr{R}(\mathbf{E}^*.\mathbf{u}). \qquad \ldots\ldots(56)$$

The problem of calculating α is thus reduced to that of calculating the electric field needed to maintain current neutrality. More precisely, if we take $\mathbf{u}$ as real, we are interested in the real part of β, defined as E_1/u, where E_1 is the component of $\mathbf{E}$ in the direction of $\mathbf{u}$.

† In all that follows we are thinking particularly of collisions between electrons and impurities, though similar behaviour is to be expected from electron–phonon collisions.

At this point the collisions between electrons and lattice come into prominence, for although they do not directly contribute to the attenuation, they do play an important role in the establishment of the electron current. We must imagine that there is a periodic electric field $\mathbf{E}$ (all field variables are to be understood as multiplied by $e^{i(\omega t - qz)}$, so that $\partial/\partial t = i\omega$, $\partial/\partial z = -iq$; the attenuation is taken as very small so that $e^{-\frac{1}{2}\alpha z}$ may be put equal to unity in the calculation of $\mathbf{E}$) associated with the lattice velocity $\mathbf{u}$, and that the electrons move through this field, acquiring momentum from it and from collisions with the moving lattice. If we confine our attention to the case of isotropic elastic scattering, we may suppose that the effect of a collision is to return the electron to some point on the Fermi surface, taking into account the fact that the Fermi surface is centred on the origin with respect to an observer moving with the lattice. This enables us to separate the problem into two parts. First we imagine the lattice at rest, and calculate the current density due to $\mathbf{E}$ and to collisions with the stationary lattice; then we put $\mathbf{E} = 0$ and calculate the current density due to collisions with the moving lattice. To take the second contribution first, it is convenient to replace the effect of collisions with the moving lattice by a fictitious electric field $\mathbf{E}_c$, by the following argument. To calculate the average displacement in reciprocal space suffered by an electron on the Fermi surface we must take a weighted average over its past history. Let us suppose that at an instant t earlier it is moving through a region where the lattice velocity is $\mathbf{u}(t)$; the chance of a collision in the interval dt leading to a change in its velocity by an amount $\mathbf{u}$ is dt/τ, and the chance that it will survive for a time t without further collision is $e^{-t/\tau}$. Thus the average additional velocity of electrons of this trajectory when $t = 0$ is

$$(1/\tau)\int_0^\infty \mathbf{u}(t)\, e^{-t/\tau}\, dt.$$

If instead of lattice movement we had an electric field $\mathbf{E}_c$ equal to $m\mathbf{u}/e\tau$ the result would be just the same. Both contributions to the current may therefore be calculated at the same time by imagining the lattice to be stationary and the field to be $\mathbf{E} + m\mathbf{u}/e\tau$.

So far what we have said applies both to transverse and to longitudinal waves. It is now convenient to treat them separately, and we take transverse waves first, as being simpler. If the lattice movement is parallel to the unit vector $\mathbf{j}$, the electronic current to be established is Neu_j. When there is no magnetic field present, the field needed to establish the current is also parallel to $\mathbf{j}$, but in a magnetic field there may be other components. Since, however, the attenuation is determined by the component E_j, we need only concern ourselves with the element $\rho_{jj}(q, \omega)$ of the resistivity tensor appropriate to a periodic field, writing

$$Ne\rho_{jj}u_j = E_j + E_c = E_j + mu_j/e\tau, \qquad \ldots\ldots(57)$$

so that

$$E_j = \frac{m}{e\tau}(\gamma_t - 1)u_j,$$

in which $\gamma_t = \sigma\rho_{jj}$ and σ is the d.c. conductivity $Ne^2\tau/m$. Hence from (56)

$$\alpha_{trans} = \frac{Nm}{Mv_{at}\tau}\mathscr{R}(\gamma_t - 1), \qquad \ldots\ldots(58)$$

in which v_{st} is the velocity of transverse waves. In the absence of a magnetic field ρ_{jj} is simply the reciprocal of the corresponding conductivity σ_{jj}, which may be calculated in the same way as was used in deriving (16). Strictly one should make allowances for the fact that the transverse field is moving with velocity v_{st}, but as for most metals this is much smaller than the Fermi velocity v_0 the error involved in supposing the field to be stationary (i.e. $\omega = 0$) is negligible. The calculation then yields the result

$$\gamma_t = \frac{\frac{3}{8}a^3}{(a^2+1)\tan^{-1}a - a}, \quad \text{where} \quad a = ql, \qquad \ldots\ldots(59)$$

which may be inserted into (58) to give α_{trans}. We shall return later to discuss this result in more detail.

For longitudinal waves it is not permissible to neglect effects arising from the movement of the electric field associated with deformation, because this field is so large (yet so phased as not to contribute to the attenuation) that effects of relative magnitude v_s/v_0 are still large enough (if correctly phased) to make major contributions to the attenuation. To substantiate this statement in more detail let us start by considering the equilibrium of a metal subjected to a static longitudinal deformation ζe^{-iqz}, which results in a periodic dilation $-iq\zeta e^{-iqz}$. In consequence the equilibrium density of electrons and the kinetic energy at the Fermi surface vary according to the laws

$$\frac{\Delta N}{N} = iq\zeta e^{-iqz},$$

$$\frac{\Delta \epsilon_0}{\epsilon_0} = \tfrac{2}{3}iq\zeta e^{-iqz},$$

since for a free electron gas $\epsilon_0 \propto N^{2/3}$. In equilibrium the total Fermi energy is constant, so that there must be developed such space charges as will maintain a periodic potential ϕ so that $\epsilon_0 + e\phi$ is independent of z;

i.e. $$e\phi = -\tfrac{2}{3}iq\zeta\epsilon_0 e^{-iqz} = -\tfrac{1}{3}iq\zeta m v_0^2 e^{-iqz}.$$

The longitudinal electric field E_0, derived from ϕ, is given by the expression

$$eE_0 = iqe\phi = \tfrac{1}{3}q^2\zeta m v_0^2 e^{-iqz}. \qquad \ldots\ldots(60)$$

An electron at the Fermi surface, moving in this statically deformed lattice, remains always on the Fermi surface, which itself varies in size from point to point; there is thus no current associated with E_0. But as soon as we allow the whole pattern of dilation and electric field to move uniformly with velocity v_s, a current is set up. To calculate this current, consider an electron, initially on the Fermi surface, moving with velocity v_0 at an angle θ to the z-axis. In traversing a distance dl it gains kinetic energy $eE_0\,dl\cos\theta$ from the field; meanwhile the field pattern has moved forwards a distance $v_s\,dl/v_0$ and the electron finds itself at a point where its potential energy has not changed by $-iqe\phi dl\cos\theta$, i.e. $-eE_0\,dl\cos\theta$, but by $-eE_0\,dl(\cos\theta - v_s/v_0)$. It has therefore departed from the local Fermi surface to the extent given by the increment of energy $eE_0 v_s\,dl/v_0$. The same effect would be produced if the field pattern were stationary and each electron were subjected to a fictitious force F, of magnitude $eE_0 v_s/v_0$, directed parallel to its motion. We may calculate the resulting current density J_t' produced by this fictitious force in the

usual way, integrating over the trajectories of all electrons at the Fermi surface. If there is no magnetic field present we easily obtain the result

$$J_i' = \tfrac{3}{4}\sigma E_0 (v_a/v_0) \int_{-1}^{1} \frac{x\,dx}{1-iax}.$$

It will be noted that this integral is purely imaginary, so that the current J_i' is in phase quadrature with E_0. It is easy enough to evaluate the integral, but a tidier result is obtained by introducing yet another fictitious field E_i', which is that electric field which would produce the same current J_i' as is produced by F (it should be noted that F acts parallel to each electron trajectory and is not therefore derivable from a potential like E_i'). On carrying out a calculation of the current which would be produced by a stationary field E_i', we derive an equation which determines E_i':

$$J_i' = \tfrac{3}{4}\sigma E_0 (v_a/v_0) \int_{-1}^{1} \frac{x\,dx}{1-iax} = \tfrac{3}{4}\sigma E_i' \int_{-1}^{1} \frac{x^2\,dx}{1-iax},$$

from which, since only the real part of the right-hand integral exists and has the same form as the imaginary part of the left-hand integral, we find

$$E_i' = ia(v_a/v_0) E_0 = i\omega\tau E_0. \qquad \ldots\ldots(61)$$

It is interesting that the same simple result is obtained when there is a magnetic field present, as is shown by Rodriguez' (1958 b) analysis, but we are unable to provide an intuitive physical reason for this. At any rate it enables us to cast the theory into the same form whether or not a magnetic field is applied. Since the lattice velocity u_i is related to displacement ζ by the equation $u_i = i\omega\zeta$, (60) may be used to write the amplitudes of the real and fictitious fields in the form

$$E_0 = -\tfrac{1}{3}iqu_i\, mv_0^2/ev_a,$$

$$E_i' = \frac{1}{3}\frac{mu_i}{e\tau}a^2.$$

The true field E_0, being in phase quadrature with u_i, contributes nothing to the attenuation and may henceforth be neglected. But E_i' is in phase with u_i, and may be very large when the electronic free path is large, since it is $\frac{1}{3}a^2$ times as great as the fictitious 'collision field' E_c.

We may now repeat for longitudinal waves the treatment of transverse waves already given, writing instead of (57)

$$Ne\rho_{ii}u_i = E_i + E_c + E_i' = E_i + \frac{mu_i}{e\tau}(1+\tfrac{1}{3}a^2), \qquad \ldots\ldots(62)$$

where E_i is the electric field which supplements E_0 and supplies as much of the required current as is not supplied by E_0 and E_i'. Since E_i' is introduced when the moving field configuration is replaced by an equivalent stationary field, ρ_{ii} is to be interpreted as the longitudinal resistivity for a stationary field of wave number q. From (56) and (62) we find

$$\alpha_{\text{long}} = \frac{Nm}{Mv_a\tau}\mathscr{R}(\gamma_1 - 1 - \tfrac{1}{3}a^2), \qquad \ldots\ldots(63)$$

in which v_{sl} is the velocity of longitudinal waves and $\gamma_l = \sigma\rho_{ii}$.† In the absence of an applied magnetic field, ρ_{ii} is the reciprocal of σ_{ii}, which is readily calculated to yield the result

$$\gamma_l = \frac{a^3}{3(a-\tan^{-1}a)}. \qquad \ldots\ldots(64)$$

It will be observed that as $a\to\infty$, $\gamma_l \to \frac{1}{3}a^2$. Thus the presence of the term $\frac{1}{3}a^2$ in (63) serves to reduce enormously the attenuation which would be observed in its absence, when the free path is very long. We saw (p. 198) that as $l\to\infty$ the conductivity associated with a periodic longitudinal field tends to zero, in contrast to that for a transverse field, and we might therefore have expected longitudinal ultrasonic waves to be highly attenuated in the range $ql \gg 1$. But in fact under these conditions the current needed for charge neutrality is almost all supplied automatically by the retardation effect which is reflected in the fictitious field E_i', and the attenuation remains finite.

In order to appreciate in more detail the behaviour of α_{trans} and α_{long} in zero magnetic field it is convenient to see the limiting forms at low and high frequencies ($ql \ll 1$ and $ql \gg 1$ respectively), as deduced from (58) and (59) and from (63) and (64):

$$\alpha_{trans} = \frac{Nmv_0q^2l}{5Mv_{st}}; \quad ql \ll 1 \qquad \ldots\ldots(65)$$

$$\frac{4Nmv_0q}{3\pi Mv_{st}}; \quad ql \gg 1 \qquad \ldots\ldots(66)$$

$$\alpha_{long} = \frac{4Nmv_0q^2l}{15Mv_{sl}}; \quad ql \ll 1 \qquad \ldots\ldots(67)$$

$$\frac{\pi Nmv_0q}{6Mv_{sl}}; \quad ql \gg 1. \qquad \ldots\ldots(68)$$

It will be seen that α_{trans} and α_{long} are very similar in behaviour, varying as q^2l when ql is small, and as q when ql is large. Since for both αl is a function of ql it is convenient to display the full range of behaviour in terms of these parameters, as in figure 38. So far as $l\alpha_{trans}$ is concerned it has been assumed that l is so long that the region in which $ql \gg 1$ is reached well before the frequency has become too high for current neutrality to be maintained. It is clear from (66) and (68) that when $ql \gg 1$, α is independent of l. This is verified experimentally, as can be seen from figure 39, with remarkable precision, which is rather surprising in view of the inadequacy of the free-electron model to describe a complex metal like indium.

The way in which α becomes independent of l recalls the anomalous skin effect, and indeed the explanation is the same in the two cases, with this difference, that with the ultrasonic wave one is in effect exploring the response of the electrons to one Fourier component of the electric field, rather than to a rather wide spectrum of components (as in the anomalous skin effect), all of which behave similarly as regards independence of l. When ql is very great the current carrying ability is

† The foregoing analysis, in terms of fictitious fields, is a not wholly successful attempt to cast the theory in a form which enables the physical content to be appreciated. More conventional analyses may be found in the references cited at the beginning of this section. In view of the assertion by several authors that the expressions for α_{trans} and α_{long}, which are the same as those derived by Pippard (1955), are in error through an inconsistency in the definition of τ, it is worth affirming categorically that this is not so ; the relaxation time which enters here is defined in the same way as that used in conductivity theory.

dominated by a small group of effective electrons moving nearly normal to **q**, whose velocity component parallel to **q** is just the velocity of sound, for these electrons move in a constant field and interact strongly with the acoustic wave. It is interesting to observe that one can reach the same conclusion by starting with the

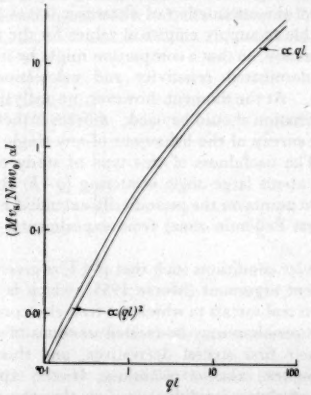


Figure 38. Theoretical attenuation curves for ultrasonic waves, shown as variation of αl with ql. The upper curve is for longitudinal, the lower for transverse, waves.

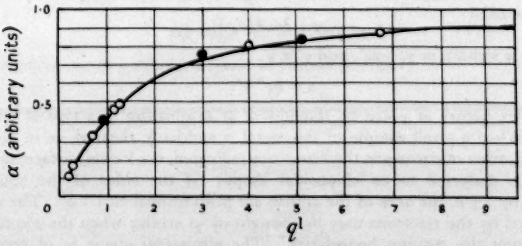


Figure 39. Experimental variation of attenuation with free path for longitudinal waves ; ○ Indium, ● Copper.
The full curve is the theoretical variation. (*After* Morse 1959.)

idea of the acoustic wave as a directed stream of phonons. The process of attenuation is then brought about by an electron–phonon collision which destroys the phonon, whose energy and momentum are taken by the electron. It is easy to show that if the electron wave number **k** is much greater than the phonon wave number **q**, so that the electron is scattered through only a small angle, the only electrons which can take part in the interactions are those which move so as to

keep in phase with the wave. There is, in fact, a very close connection between the semi-classical treatment of ultrasonic absorption given here and the quantal (or semi-quantal) treatment of electron–phonon scattering which is normally applied to conduction problems in metals. In the low frequency limit ($q \ll k$ for an electron on the Fermi surface) they give the same answer, provided $ql \gg 1$.† This suggests that a detailed survey of the attenuation of ultrasonic waves as a function of crystal orientation might be able to supply empirical values for the matrix elements of the electron–phonon interaction, so that a comparison might be made between the high-temperature phonon-dominated resistivity and calculations based on realistic interaction parameters. At the moment, however, no analysis has been made of the way in which the information should be used; moreover there has been no attempt yet to make a detailed survey of the behaviour of any single metal as a function of crystal orientation. The usefulness of this type of study really depends on how much can be inferred about large-angle scattering ($q \sim k$) and Umklapp processes (scattering between two points on the periodically extended Fermi surface, of which only one lies in the first Brillouin zone) from experimental information on small-angle scattering.

The attenuation under conditions such that $ql \ll 1$, as given by (65) and (67), may be derived by a different argument (Morse 1955), which is of value in discussing what can be expected in real metals to which the free-electron model is inapplicable. When $ql \ll 1$ the whole problem may be treated in terms of the local values of the field variables and their first spatial derivatives, and this is equivalent to the calculation of a (complex) elastic modulus. If the appropriate compliance coefficient (reciprocal of elastic modulus) is χ so that the wave velocity may be written as $(\chi M)^{-\frac{1}{2}}$, a complex χ implies a complex v_s and hence attenuation, the attenuation constant as earlier defined being related to the imaginary part of v_s^{-1} by the expression

$$\alpha = -2\omega \mathscr{I}(v_s^{-1}),$$

so that if χ is written as $\chi' - i\chi''$, and $\chi'' \ll \chi'$,

$$\alpha = q\chi''/\chi'. \qquad \ldots\ldots(69)$$

The complex nature of χ can be thought of as originating in a typical relaxation process. When a small sample of the metal is suddenly sheared, as in figure 40, by the application of stresses in the directions indicated, the Fermi surface, originally spherical, is distorted to an ellipsoidal shape; if the sides of the square are multiplied by $1 \pm \delta$, the axes of the ellipse are proportional to $1 \mp \delta$. The changes of $\mathbf{k}$ suffered by the electrons may be thought of as arising when the electrons are reflected from the moving boundaries. The ellipsoidal shape is of course not the equilibrium form of the Fermi surface, and electron–lattice collisions lead to a re-establishment of the spherical form with characteristic relaxation time τ. Now to produce the deformation the stresses P have in the main to overcome the rigidity of the lattice, but there is a small additional dynamical term due to the pressure created by electronic collisions with the boundary. In equilibrium this takes the well-known value $\frac{1}{3}N\overline{mv^2}$, which for a degenerate Fermi gas of free electrons is

† If ql is not much greater than 1 a quantal theory is lacking. It may be expected from the semi-classical theory that free path effects will be important in electron–phonon interactions if $ql \sim 1$: this matter has been discussed briefly (Pippard 1957 c), and there is some experimental evidence supporting the conclusions. A more rigorous quantal treatment is most desirable.

$\frac{1}{5}Nmv_0^2$; but when the Fermi surface is deformed into an ellipsoid the velocity components must be appropriately scaled, so that the pressures on different surfaces are unequal. In figure 40 the pressure on the boundaries normal to x and y are $\frac{1}{5}Nmv_0^2(1-2\delta)$ and $\frac{1}{5}Nmv_0^2(1+2\delta)$ respectively. The external stresses required to produce the initial deformation are thus rather larger than are required to maintain it once relaxation has occurred, the difference being $\frac{2}{5}Nmv_0^2\delta$. Since

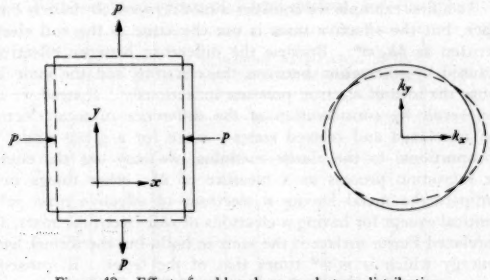


Figure 40. Effect of sudden shear on electron distribution.

the angle of shear in the present case is 2δ, the shear compliance is $2\delta/P$. If then we write χ_0 for the long-term compliance and χ_∞ for the short-term compliance, we have that

$$P = 2\delta/\chi_\infty \quad \text{and} \quad P - \tfrac{2}{5}Nmv_0^2\delta = 2\delta/\chi_0,$$

so that

$$\Delta\chi \equiv \chi_0 - \chi_\infty = \tfrac{1}{5}Nmv_0^2\chi_0^2. \qquad \ldots\ldots(70)$$

We now make use of a standard result of relaxation theory

$$\chi(\omega) = \chi_\infty + \frac{\Delta\chi}{1+i\omega\tau},$$

so that

$$\chi'(\omega) = \chi_\infty + \frac{\Delta\chi}{1+\omega^2\tau^2}$$

and

$$\chi''(\omega) = \frac{\omega\tau\Delta\chi}{1+\omega^2\tau^2}.$$

Since we are limited to the case when $ql \ll 1$, it follows that $\omega\tau$, which equals $v_s ql/v_0$, is extremely small, so that we may write $\chi'(\omega) = \chi_0$ and $\chi''(\omega) = \omega\tau\Delta\chi$. Hence from (69) and (70), for a transverse wave

$$\alpha_{\text{trans}} = \tfrac{1}{5}Nmv_0^2\chi_0 q\omega\tau = \frac{Nmv_0 q^2 l}{5Mv_{sl}},$$

exactly as in (65). We may also reproduce the result for longitudinal waves by recalling that the appropriate elastic modulus is $K+\frac{4}{3}n$, where K is the bulk modulus and n the shear modulus. For a free-electron gas there is no imaginary

part of K, since uniform compression does not destroy equilibrium. The attenuation is thus all attributable to the shear modulus, and α_{long} is obtained from α_{trans} by multiplying by $\frac{4}{3}$ and replacing v_{st} by v_{sl}, as in (67).

By interpreting the attenuation in terms of relaxation processes contributing to the elastic compliance we may investigate qualitatively the types of contribution which may be expected in real metals, where the free electron model is inapplicable, and for which the fuller treatment, valid for arbitrary ql, is complicated and full of pitfalls. As a first example we consider a model in which there is one spherical Fermi surface, but the effective mass is not the same as the real electron mass, i.e. v_0 is written as $\hbar k_0/m^*$. Because the difference between effective and real masses is caused by interaction between the electrons and the ionic lattice it is not safe to use the idea of electron pressure incautiously. Instead we may derive the required result by consideration of the difference in total electron energy between the unrelaxed and relaxed states; since for a given strain the elastic energy is proportional to the elastic modulus, we may use the energy lost as heat in the relaxation process as a measure of $\Delta\chi$, other things being equal. Thus to compare the metal having n electrons of effective mass m^* with one which is identical except for having n electrons of real electronic mass, for a given shear the unrelaxed Fermi surface is the same in both, but the former has an excess electronic energy which is m/m^* times that of the latter. In consequence the attenuation constant, for a given value of τ, is m/m^* times that in the free-electron metal. The results expressed by (65) and (67) represent this more general case if m is replaced by m^*.

It would be unwise to infer from this that the general expressions for α can be so simply modified to allow for departures from the free electron model. Indeed it seems likely that the method of analysis used here is not applicable in general, since there is not the possibility of a unique division of the system into conduction electrons and lattice. A start towards the solution of the problem for arbitrary Fermi surfaces has been made by Blount (1959), which is sufficient to indicate how further progress may be made, but which has not yet yielded any results of wide applicability. We therefore revert to our discussion of such cases as can be treated as local relaxation phenomena.

The process we have analysed, relaxation following a shear deformation, is exactly analogous to viscosity in a liquid, and in fact one of the first explanations (Mason 1955) of the observed high attenuation of ultrasonics in pure metals at low temperatures was in terms of the viscosity of the electron gas. Since the viscosity of a gas is proportional to the free path, we expect as l increases at low temperatures to find a greater viscous damping of the lattice oscillations. This expectation is of course also expressed in (65) and (67) when $ql \ll 1$, and is verified by observation, as shown in figure 41. This figure also shows how the attenuation drops when the metal becomes superconducting. From observations (Bömmel 1955) on well-prepared single crystals it has been made fairly certain that the electronic contribution to α vanishes as the temperature goes to zero. This result has a certain value in the study of the non-superconducting state of the metal, for it enables extraneous sources of attenuation to be allowed for. Thus α does not drop to zero in the superconducting state of a polycrystalline sample, presumably because of scattering at grain boundaries which is unaffected by the superconducting transition. If

one may assume that the residual, non-electronic attenuation is temperature-independent in the range of interest, the electronic part may be determined accurately by subtraction. We shall say nothing more of the behaviour in the superconducting state, which is of great interest but not obviously relevant to the present study.

Once we leave the highly idealized models of metals with one spherical Fermi surface, other effects can come in and make important contributions to the attenuation. With one non-spherical Fermi surface, for example, it is possible

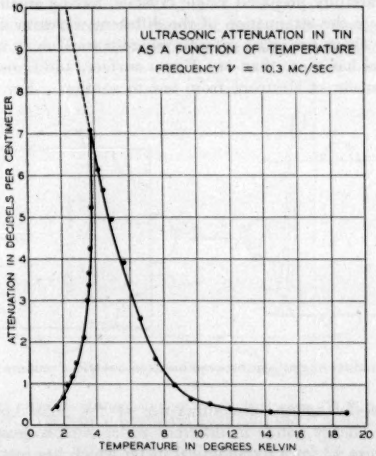


Figure 41. Temperature variation of attenuation of longitudinal wave (10·3 Mc/s) in a very pure single crystal of tin. The superconducting transition occurs at 3·7°K. (*After* Bömmel 1955.)

to have losses associated with pure compression as well as with pure shear, in hydrodynamical terms second as well as first viscosity. It is only necessary that the equilibrium form of the Fermi surface shall depend on the volume of the metal, so that a sudden uniform compression, which produces a uniform expansion in reciprocal space, leaves the Fermi surface in a non-equilibrium state. One would not expect this effect to be important in a metal whose Fermi surface was nearly spherical, but it might well be comparable to the effect of shear in a metal whose Fermi surface is very close to, or just touches, the zone boundary; for the energy gap at the boundary and consequently the degree of contact are probably sensitive functions of the atomic volume. In principle it should be possible to estimate the sensitivity of the Fermi surface to volume changes by measuring the imaginary parts of the bulk modulus, and hence to discover whether there is likely to be contact with the zone boundary. Measurements of the attenuation of longitudinal and transverse waves on the same sample, under conditions such that $ql \ll 1$, appear at first sight promising, but there are difficulties. Since all metals

are elastically anisotropic, the simple concepts of bulk and shear moduli are inapplicable except to a finely polycrystalline sample whose crystallites are small compared to the wavelength. If such samples are used there will probably be grain-boundary attenuation as well as electronic attenuation, and the only method known at present to estimate this is to use the superconducting behaviour already discussed.† Unfortunately the group I metals for which this study would be particularly interesting, being the only metals in which the Fermi surface is likely to have but one sheet, do not become superconducting. It would therefore be necessary to use carefully prepared single crystals, having small residual attenuation, and to measure the attenuation of the different waves in different directions chosen so that the shear processes in each were identical.

When the metal has more than one Fermi surface, additional attenuation may result from the transfer of electrons from one to another. For example, uniform

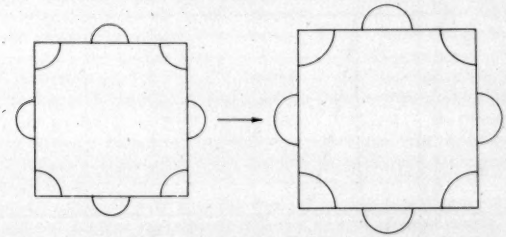


Figure 42. Transfer of electrons between bands as a result of uniform compression.

compression alters the energy discontinuities at the zone boundary and may therefore shift the energy bands in different zones with respect to one another, as illustrated in figure 42 for a hypothetical metal which has just enough electrons to fill the zone, but a small energy overlap so that the Fermi surface occupies two zones. In the cases we have mentioned the calculation of the attenuation constant when $ql \ll 1$ reduces to a calculation of the energy difference between the unrelaxed and relaxed electron distributions, combined with a knowledge of the relaxation time for the appropriate transition. If the shape of the energy surfaces is highly dependent upon deformation, and especially if the effective mass is small so that the transfer of a small number of electrons involves a considerable energy change, the metal may show an abnormally high attenuation in comparison with the free-electron model. Such is presumably the explanation of the fact that α is quite readily measurable in bismuth (Reneker 1959), although the effective number of conduction electrons is extremely small.

When ql is greater than unity, the current which balances the effective lattice current is not carried by all electrons, but more particularly by those whose velocity

† It might be possible to correlate the temperature variation of attenuation with that of electrical resistance and so detect the temperature-invariant part of the attenuation. It is not certain, however, that the phonon-controlled mean free path is the same in the two phenomena. Presumably a careful study of superconducting metals would enable this point to be investigated, so that ultimately some use could be made of polycrystalline non-superconductors.

component in the direction of wave propagation equals the wave velocity. One may expect that the overall sensitivity of the Fermi surface to deformation is still a measure of the effective lattice current, which is to say that one may expect (58) and (63) still to describe qualitatively the variation of α with ql. But there will be quantitative differences depending on the properties of the electrons on the effective zone in relation to the average over the whole Fermi surface. A complete theory has not been worked out, and until it has one should be cautious about the use of the free-electron theory for the quantitative interpretation of results on real metals.

We turn now to the effect of a magnetic field on the attenuation, considering first a free-electron model for which the calculations are straightforward in principle,

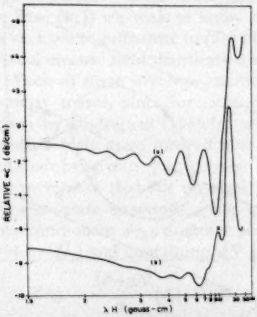


Figure 43. Oscillatory variation of attenuation of longitudinal waves (75 Mc/s) in a very pure single crystal of copper. The wave is propagated along [001] and H is applied along [100]. (*After* Morse and Gavenda 1959.)

even though they may involve a considerable amount of numerical work to evaluate the answer. The general results expressed by (58) and (63) are still valid, and the problem reduces to a calculation of γ. What is found in practice (Bömmel 1955) is that when $ql \gg 1$ the attenuation is an oscillatory function of field strength, as shown in figure 43. A naïve interpretation of this, which appears to agree with the experiments as regards the field strength involved, relates the variations of α to the relative sizes of wavelength and orbit diameter. For example, in the simplest case (**j**, **j**) the orbits and electric fields associated with the wave are represented in figure 44 ; **E** and **u** are parallel and ρ_{jj} is the reciprocal of the conductivity in this direction. If the field varies as $\mathbf{E}\,e^{i(\omega t - qz)}$ and the orbit, centred at z_0, has radius R, the average field strength around the orbit is $\mathbf{E}(z_0)J_0(qR)$. This is an oscillatory function of qR whose absolute amplitude has maxima whenever the orbit diameter is very nearly $(n+\frac{1}{4})$ wavelengths. When we take into account all electrons at one point having a given orbit diameter we must allow for a distribution of centres, which leads to the average field experienced being $E(z_0)\,[J_0(qR)]^2$. If all the electrons had the same orbit diameter their effective conductivity would contain the term $[J_0(qR)]^2$, and so γ and α would be oscillatory, the maxima of

$[J_0(qR)]^2$ corresponding to maximum conductivity and minimum attenuation. Since $R \propto H^{-1}$ we might expect to find α an oscillatory function of H^{-1} with regular period. This is similar to the de Haas–van Alphen effect and cyclotron resonance, but here the effect involves the spatial extensions of orbit and wave, not their periods as in cyclotron resonance. As a result the frequency at which the magneto-acoustic oscillations should occur is of the order of a hundred times less than the cyclotron resonance frequency in the same magnetic field, the ratio being $0{\cdot}26v_0/v_*$. This agrees well with what is found in practice (Morse, Bohm and Gavenda 1958), and it is disturbing therefore to discover that a full working of the theory destroys the predicted effect in this particular case. If the Fermi surface is spherical different electrons on the surface have different orbit radii,

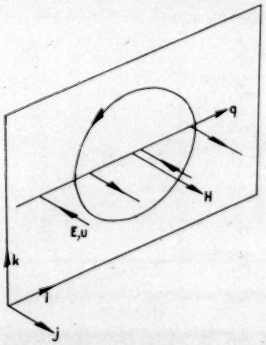


Figure 44. Illustrating fields and an electron orbit in the case (j, j).

and it is necessary to integrate over all parts of the Fermi surface. The final result is that for a sphere of radius k_0, when $ql \gg 1$,

$$1/\gamma = 3\int_0^1 [J_0(mx)]^2 x\sqrt{(1-x^2)}\,dx, \quad \text{where} \quad m = \hbar k_0 q/eH. \qquad \ldots\ldots(71)$$

The meaning of this expression is clear; $3x\sqrt{(1-x^2)}\,dx$ is the fraction of electrons lying within a cylindrical shell of radius $k_0 x$ and thickness $k_0\,dx$, having its axis parallel to **H**. When **E** is parallel to **H**, as here, this is the weighting factor to be ascribed to the orbits of radius $\hbar k_0 x/eH$. The vanishing of the weighting factor at the upper limit of integration corresponds to the fact that with this geometry the electrons of largest orbit make a vanishing contribution to the current, for their velocity is normal to the electric field. There is thus no especial orbit size which can dominate the conductivity, and it is not surprising to find on carrying out the integration of (71) that γ is a monotonic function of H. If the Fermi surface is not spherical it may be possible for one particular orbit size to be sufficiently dominant to produce oscillations in γ; for example a cylindrical Fermi surface, oriented so that its axis is not parallel or normal to **H**, will have

all electrons on the surface executing the same elliptic orbit, and very marked oscillations can be expected. But in less extreme cases one can hardly hope for anything better than a broad spectrum of orbit sizes, so that even if any oscillations are seen they will be confined to one or two cycles only, higher harmonics (orbit diameters of several wavelengths) being smeared out. If this should be the explanation of the oscillations observed, it is of interest as showing the phenomenon to be one which is dependent on the Fermi surface departing from spherical shape in a particularly favourable way. But there are other possibilities of explanation, which we defer until we have discussed the other transverse field geometries (**k**, **j**) and (**i**, **j**).

Here the situation is a little more complex, since it is not sufficient to have **E** and **u** parallel to get **J** and **u** parallel; both transverse and longitudinal electric fields are needed. In the case (**k**, **j**) we wish to know what field E_k is needed to produce a given J_k, that is we have to substitute in (58) the component of resistivity ρ_{kk} appropriate to a sinusoidal electric field distribution; in the case (**i**, **j**) we have to substitute ρ_{ii} in (63). Each of these involves calculating the components σ_{kk}, σ_{ii} and σ_{ik} of the conductivity tensor, since for example $\rho_{kk} = \sigma_{ii}/(\sigma_{ii}\sigma_{kk} + \sigma_{ik}^2)$. Expressions have been given by Rodriguez (1958 b), which although correct are unfortunately written in such a form as to mislead him into making a number of erroneous statements about their behaviour, particularly that they are not oscillatory (it is perhaps unnecessary to remark that his physical argument to explain why they are not oscillatory is also quite unsound). For example in the interesting limit $ql \gg 1$ the conductivity component $d\sigma_{kk}$ due to a plane section of the electron assembly, cut normal to **H** ($= H_j$) and containing dN electrons per unit volume, takes the form

$$d\sigma_{kk} = \frac{2e^2\tau dN}{m}[J_1(qR)]^2, \qquad \ldots\ldots(72)$$

which is certainly oscillatory, as are σ_{ik} and σ_{ii} also. In contrast to the case (**j**, **j**), the central section of the Fermi sphere now has considerable weight, for the displacement of the surface of the section carries, as it were, all the inner electrons with it; alternatively expressed, the central electrons have their velocities in the plane normal to **H** which also contains **E**, and are therefore efficient conductors. In fact when we come to integrate over all sections the situation is like what we found in cyclotron resonance, that the extremal section dominates the average, and the components of the conductivity oscillate with a periodicity characteristic of the extremal section. We thus expect to find an even periodicity when α is plotted against H^{-1}, the separation of neighbouring maxima being $\pi e/\hbar q k_0$. The curve shown in figure 43 exhibits this clearly; from the number of periods visible one may feel sure that there is no averaging of frequencies involved, but that one group of orbits is dominating the behaviour. It is therefore rather disturbing to find that the value of k_0 deduced from this curve is only 0·8 of what is expected; the discrepancy is large enough to need explanation.

It is possible that the explanation lies in the non-spherical nature of the Fermi surface of copper—not that it is likely to have any dimension as small as 0·8 times the diameter of the free-electron sphere (see p. 258), but it may be that the theory of the effect in non-spherical surfaces will reveal some unconsidered remedy for the discrepancy. This is not however a very hopeful suggestion; one would

surely expect the periodicity to be governed by the full extension of the orbit in the direction of **q**, since it is at the extremes of the orbit that the electron spends a significant time in a constant field. If this difficulty can be cleared up the study of magneto-acoustic oscillations† should prove a useful method of determining specific dimensions of the Fermi surface. It is therefore worth examining the conditions which must be satisfied if the oscillations are to be observed. The electronic free path must be long enough for an orbit to be completed between collisions; one may guess a collision damping factor of the form $e^{-2\pi R/l}$, so that l must be at least as big as the orbit diameter, or perhaps three or four wavelengths, if several cycles of the oscillation are to be observed. For a sound velocity of 2×10^5 cm sec^{-1} and a frequency of 100 Mc/s this implies that l must approach 10^{-2} cm, which is at present achievable only with very few metals. The magnetic fields required are quite modest, of the order of 10 kG, so that the main limitation of this method is the purity of the sample.

The possibility of oscillatory behaviour in the cases (**k**, **j**) and (**i**, **j**) makes one suspicious about the explanation of the oscillations in (**j**, **j**) tentatively advanced above. For the elastic anisotropy of the metals used implies that only along very special directions of high symmetry will the wave be purely transverse or longitudinal. In a polycrystal most crystallites will have electric fields with components not parallel to **H**, and it is possible that these are responsible for the oscillatory behaviour. One cannot be sure of this, but it could be checked by a study of single crystals of copper, as pure as those used by Morse to obtain figure 43, to see whether the oscillations only appear when **q** does not lie along an axis of high symmetry.

One last point to be made in connection with the transverse field effects concerns the limiting behaviour in high magnetic fields, when $qR\ll1$. The different cases have theoretically different limits. For example in (**j**, **j**) the coiling of the orbits into a small space allows lattice collisions to take over entirely the business of establishing the electronic current; γ tends to unity, and α to zero, as H^{-2}. In the other two cases the combined effect of lattice collisions and magnetic field is to establish a current at right angles to the required direction, and electric fields are still needed. This does not prevent the attenuation going to zero as H^{-2} in the other transverse case (**k**, **j**), but in the longitudinal case (**i**, **j**) α tends to the limit $\frac{1}{15}(nmv_0q^2l/Mv_{\rm el})$. This has the same form as in zero field, when $ql\ll1$ (see (67)), but with one quarter of the coefficient. Since for observing the oscillations ql must be much greater than unity, the limiting value of α will be several times the value in zero field. It should of course be remembered that this analysis of the limiting behaviour is only valid for a spherical Fermi surface; magneto-resistance may considerably alter the behaviour in high fields.

The last magneto-acoustic effect to be treated is the case (**j**, **i**) discussed in some detail by Kjeldaas (1959). It is convenient here to think of the transverse wave as circularly polarized, setting up a circularly polarized electric field. Since **H** is parallel to **q** the electrons perform helical paths about **q** as axis, and there is one particular group which have the correct component of velocity parallel to **q** to rotate in their orbits at just the same rate as the field in which they find

† In a short note (Pippard 1957 b) proposing this method of studying metals the effect was referred to as magneto-acoustic *resonance*. This is not a resonance phenomenon in the accepted sense and the term should be dropped. It is as well to point out also that in this note, through faulty reasoning, the condition of minimum attenuation was stated to be one of maximum attenuation.

themselves. These electrons in synchronism with the field are particularly good conductors and when $ql \gg 1$ make a dominant contribution to the real part of the effective conductivity. The phenomenon can be thought of as a cyclotron resonance between the electrons and the ultrasonic wave, but the frequency of the latter is enormously shifted by the Doppler effect of the electronic motion. We shall analyse the problem for a very special simple model in which the Fermi surface has axial symmetry about **q** (see figure 45). An electron at O has cyclotron

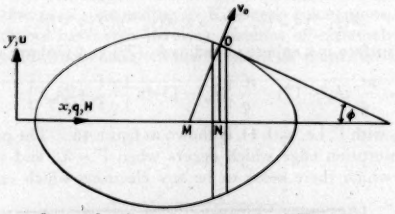


Figure 45. Illustrating the Kjeldaas (1959) effect.

frequency $\omega_c = Hev_0 \cos\phi/(\hbar k_y)$, where k_y is the radius ON, and is moving in the x-direction, parallel to **q**, with velocity $v_0 \sin\phi$. The Doppler-shifted frequency ω' which it experiences is $q(v_0 \sin\phi - v_s)$, and the relative frequency is thus given by the expression

$$\Delta\omega \equiv \omega_c - \omega' = qv_0 \sin\phi(\Gamma/S - 1 + v_s \operatorname{cosec}\phi/v_0), \qquad \ldots\ldots(73)$$

where $\Gamma = eH/(\hbar q)$ and $S = k_y \tan\phi = MN$.

If we neglect the term $v_s \operatorname{cosec}\phi/v_0$, which is equivalent to assuming the wave to be at rest, the resonance condition is seen to be that $\Gamma = S$, S being the sub-normal of the curve at O. As **H** and hence Γ change, different parts of the Fermi surface come into resonance. The calculation of the effective conductivity σ_{eff} is straightforward, since all that is necessary is to forget about **H** and instead to imagine each electron in a uniform electric field which rotates with frequency $\Delta\omega$. A short calculation then shows that

$$1/\gamma_t = \sigma_{\text{eff}}/\sigma = C \int \frac{lk_y \cos\phi}{1 + i\Delta\omega . \tau}\, dk_x, \qquad \ldots\ldots(74)$$

where

$$1/C = \int lk_y \cos\phi \, dk_x,$$

$\Delta\omega$ being, of course, a function of k_x, given by (73). We now consider what happens when l and τ are very large. The only significant contribution to the real part of γ_t^{-1} comes from the region around the resonance point where $\Delta\omega = 0$, the rest of the Fermi surface being responsible for the imaginary part of γ_t^{-1}. In evaluating $\mathscr{R}(\gamma_t^{-1})$ we may therefore take $lk_y \cos\phi$ in the integrand as constant, and after a certain amount of manipulation obtain the result

$$\mathscr{R}(\gamma_t^{-1}) = \frac{\pi C}{2q} \left| \frac{d(k_y^4)}{d(S^2)} \right|_{S=\Gamma}, \qquad \ldots\ldots(75)$$

while
$$\mathscr{I}(\gamma_t^{-1}) = -\frac{iC}{2q}\mathscr{P}\int\frac{d(k_y^4)}{d(S^2)}\cdot\frac{dS}{S-\Gamma}, \qquad \ldots\ldots(76)$$

in which $\mathscr{P}$ means that the principal value of the integral is to be taken. It will be observed that except for the field-independent C both parts of γ_t^{-1} are determined by the geometry of the Fermi surface. Now from (58) α depends on $\mathscr{R}(\gamma_t - 1)$, and both $\mathscr{R}(\gamma_t^{-1})$ and $\mathscr{I}(\gamma_t^{-1})$ enter into this. In particular

if
$$\mathscr{R}(\gamma_t^{-1}) \gg \mathscr{I}(\gamma_t^{-1}), \quad \alpha \propto 1/\mathscr{R}(\gamma_t^{-1})$$

if
$$\mathscr{R}(\gamma_t^{-1}) \ll \mathscr{I}(\gamma_t^{-1}), \quad \alpha \propto \mathscr{R}(\gamma_t^{-1})/[\mathscr{I}(\gamma_t^{-1})]^2.$$

If the Fermi surface is a sphere of radius k_0 (75) and (76) are readily evaluated to give
$$\gamma_t^{-1} = \frac{\pi C}{q}(k_0^2 - \Gamma^2) - \frac{iC}{q}\left\{(k_0^2 - \Gamma^2)\ln\frac{1+\Gamma}{1-\Gamma} + 2k_0\Gamma\right\}. \qquad \ldots\ldots(77)$$

The variation of α with Γ, i.e. with **H**, is shown in figure 46. The point of particular interest is the absorption edge which occurs when $\Gamma = k_0$, and which marks the field strength at which there cease to be any electrons which can resonate with

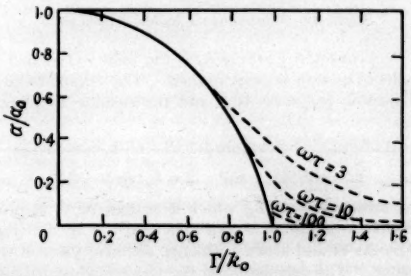


Figure 46. Variation of attenuation with field for various values of $\omega\tau$. (*After* Kjeldaas 1959).

the ultrasonic field. Such an edge may be expected to occur with any elliptic limiting point (in the nomenclature which Azbel' and Kaner use for cyclotron resonance), and it is easy to show that, whether or not the surface is axially symmetric, the value of Γ at the absorption edge is just $K^{-\frac{1}{2}}$, when K is the Gaussian curvature at the limiting point.

At first sight this elegant result appears to offer an attractive opportunity for analysing the Fermi surface, particularly of those metals for which it is difficult to prepare clean surfaces for the study of the anomalous skin effect, or those whose mosaic structure renders them unfit for de Haas–van Alphen studies. But certain reservations must be made. Even if we leave aside the difficulty of obtaining circularly polarized transverse waves,† there still remains the fact that in most

† If linearly polarized waves are used, as is experimentally much easier, the theory is hardly altered, as the wave can be analysed into two circularly polarized waves of opposite sense, which resonate with almost equivalent regions on opposite sides of the Fermi surface. There is, however, a small difference in velocity of the two waves, so that a rotation of polarization of the plane wave is to be expected. This may prove very annoying in practice.

directions of propagation the waves will not be purely transverse. It may be that a longitudinal component will not affect the form of the absorption edge, but this needs careful analysis. Along directions of high symmetry in cubic metals (e.g. (100), (110) and (111)) the present analysis should be valid, and in some cases determination of the curvature of the limiting points in these directions might almost entirely fix the form of the Fermi surface, if there is but one sheet as in the alkali and noble metals. It must also be remembered in making a provisional assessment of the method that the sharpness of the absorption edge is a consequence of a very high value of ql; according to Kjeldaas's calculations, a value of 50 is desirable, and this we have seen involves material of extremely high purity, as well as high acoustic frequencies. Finally we must point out the possibility of

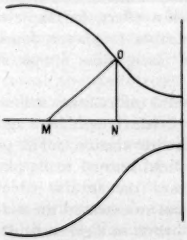


Figure 47. Maximal sub-normal when Fermi surface has no limiting point.

ambiguity in interpretation of the results. If the Fermi surface has no limiting point, because of contact with the zone boundary, it will have another region at which S is maximal (figure 47), and this will give rise to an absorption edge. The behaviour of γ_t^{-1} here is quite different from that at a limiting point, for the real part becomes infinite as $(S_{\max}-\Gamma)^{-1/2}$ while the imaginary part tends to become rather small. The attenuation goes to zero at the absorption edge, but as $(S_{\max}-\Gamma)^{1/2}$, i.e. with a vertical tangent, rather than as $(S_{\max}-\Gamma)$, with a finite slope, as it does if there is a limiting point. In principle the two types of edge can be discriminated, but one might have difficulty unless ql were very high indeed. Nevertheless this experiment promises to have considerable interest, and there is quite a good chance that it will yield information of real value.

4.6. *Size Effects*

It is possible, though difficult, to prepare a thin film or wire in a sufficiently pure state that at low temperatures by far the most potent scattering mechanism is by collision of the electrons with the surface of the sample. Measurement of the conductivity of the film or wire, combined with some assumption about the scattering properties of the surface, enable an estimate to be made in principle of the conductivity of the metal for a known free path (a calculable function now of the direction of the electronic trajectory). In a film for example the dominant electrons are those moving parallel to the surfaces, and the information which might be extracted is to all intents the same as is yielded by the anomalous skin

effect. It seems, however, that the latter is a much easier and more reliable technique, having the advantages that the exact nature of the surface scattering is of only minor importance, and the preparation of clean samples of any crystal orientation is not too difficult. With the film, on the other hand, only certain orientations of single crystals can be made, by epitaxial growth, and it is extremely hard to ensure that they shall be uncontaminated by residual gas in the evaporating oven. Moreover the films made in this way are very liable to vary in thickness and even to have holes in them which greatly obscure the interpretation of the measured resistance. Thin monocrystalline wires are more readily prepared, as whiskers, and may be obtained very pure so that surface scattering is dominant. But they also are limited in the range of orientations possible, and in addition do not yield the specific information about the electronic structure that is yielded by films and the anomalous skin effect, for the electrons moving nearly parallel to the axis are too few in number to play a dominant role. It does not seem therefore that the prospect of using these simple size effects as analytic tools is very promising.

There is more hope that useful information will come out of such measurements in the presence of a magnetic field, though here again the limitations are such as to make it no more than an outside chance. The periodic oscillations of conductivity of a film in a magnetic field normal to its plane, predicted by Sondheimer (1950) but never observed, and the similar effect observed by Babiskin and Siebenmann (1957) in a wire, but not treated theoretically in detail, might possibly yield data on the extension of orbits in a given field. It is rather disheartening to find that even for a free-electron metal with surface scattering completely dominant the predicted oscillations are so highly damped as to be observable for only two or three cycles, for this suggests that there is no one group of electrons which can govern the periodicity as in some of the other phenomena we have analysed. The interpretation of the data for a real metal is therefore likely to be difficult unless the electronic structure is already known.

Finally we may mention the proposal by Azbel' (1958) to study the Schubnikov-like oscillations of the skin resistance (see p. 236) in a rather thick monocrystalline film. He points out that no oscillations will be observed until the field is high enough for the electronic orbit to lie completely within the film, and that the field strength at which any particular oscillation appears determines one dimension of the Fermi surface. We shall not enter into this proposal any further here; a full discussion may well await the performance of the experiment. As this involves the preparation of a thick film ($> 10^{-4}$ cm) so perfectly crystalline as to show the oscillations, and so pure that the electron can complete an orbit entirely within the film, we believe that a revolutionary advance in the art of thin films will be necessary before it is worth trying.

§ 5. Applications

We shall not attempt to give a complete survey of the applications made up to the present of the methods we have discussed. It seems more profitable to take the view that these studies are still in their infancy, and to describe a few examples which illustrate the scope and limitations of the methods. We therefore pass over without comment a large amount of work on the de Haas–van Alphen

effect in moderate fields, which has shown the presence in many metals of small inclusions of holes or electrons; a full discussion has been given by Shoenberg (1957). It is not perhaps so justifiable to let pass without discussion a great diversity of valuable work on bismuth, graphite and other semi-metals, in which the Fermi surfaces are all small in extent, and the revealing properties, being capable of examination in moderate magnetic fields, have been studied in some detail. Our excuse for this neglect is that these substances form a special class of their own, and deserve an article to themselves if all that is known about them is to be placed in proper perspective.† By adopting this cavalier attitude we have narrowed down the field sufficiently to permit what is left to be examined at leisure, and we shall concentrate on results obtained for copper and lead.

The analysis of the Fermi surface of copper (Pippard 1957 a) illustrates both the strength and the weakness of the anomalous skin effect for this work. The surface resistance varies in a complicated manner with crystal orientation, the extreme variation amounting to nearly a factor of two. Only moderate precision (about 3%) was needed to reveal the anisotropy of resistance with sufficient accuracy to make detailed analysis worth the considerable effort. There is no need to recapitulate the arguments and numerical procedures which led to a single closed surface whose variations of curvature were able to explain the observations. The surface is derivable from a sphere by pulling it out along the (111) directions and slightly denting it in the (110) directions, the ratio of maximum to minimum radius being 1·16. Although the variations of radius are so small, the high (cubic) symmetry of the surface makes them sufficient to introduce slight concavities, so that there are regions of very large radius of curvature contributing strongly to the effective conductivity. This is how the large anisotropy of resistance comes about, and this is the strength of the method, that small variations of radius can lead to large variations of resistance. The shape of the Fermi surface is thus rather precisely determined, particularly around the flattened parts. The absolute size of the surface can be found no more accurately than the probable error of the readings, and although this would not be of great importance in many cases, it happens that copper is not one of these. The measurements suggest that the surface is of such a size as to hold, within a few per cent, one electron per atom. If we were genuinely dealing with independent electrons we should feel no hesitation about taking this result as confirmation that only the valence electrons occupy unfilled Brillouin zones, and scaling the surface so that it was precisely the size to hold one electron per atom, that is, half the volume of the Brillouin zone. But the quasi-particles are not independent electrons, but only something like electrons, and it is not obvious that the Fermi surface of the quasi-particles should contain as many states as there are electrons. Plausible arguments may be developed in support of this view, but they are not altogether convincing, and although we shall adopt it as the most likely hypothesis it remains a weakness in the chain of reasoning. When the derived surface is scaled to occupy half the zone it is found that it will not quite go in, the (111) extensions causing overlap beyond the zone boundaries. Now it is just at the points of maximum extension that the shape is least securely

† The following papers give some idea of what has been done in this field :
For bismuth ; Abeles and Meiboom (1956), Heine (1956 b).
For graphite ; Soule (1958 b), McClure (1958).

determined, for the radius of curvature is small here, and the contribution to the conductivity correspondingly small. By means of crude arguments it may be contended that the most plausible modification to make the surface fit into the zone is to have positive contact with the boundary, as shown in figure 48, and this is what is proposed as the Fermi surface in copper. It should be emphasized that the extent of the contact, its existence even, is conjectural, but it is almost certain from the data that the deformation of the surface from spherical form is of the type and roughly of the magnitude shown here ; that is, it is pulled out towards the (111) faces, and either touches them over a small area or just fails to touch them. The insensitivity of the experimental results to the form of a critical region like this is a real weakness of the method.

Recently the de Haas–van Alphen effect has been discovered in copper whiskers (Shoenberg 1959), and although the investigation is far from complete it is worth

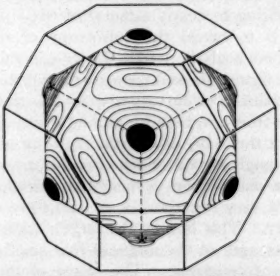


Figure 48. The Fermi surface in copper (redrawn from Pippard 1957 a). The dark patches mark contacts with the zone boundaries.

seeing what might be expected if the proposed Fermi surface is correct. For a considerable range of orientations of the magnetic field around the (100) axes (within a cone of semi-angle about 30°) and around the (111) axes (semi-angle about 15°), the central section normal to the field does not intersect the contact regions, and the section is bounded by a closed electron orbit whose area is nearly equal to that of a free-electron sphere. One would expect the period of the oscillations to vary only by a few per cent. The results up to now, of which figure 49 is typical, indicate that for these orientations only one period is present, and this is nearly independent of orientation, but the oscillations have not been followed more than 8–10° from either axis as the amplitude falls off rapidly. This may be an effect of mosaic structure, which is less deleterious when the crystal is oriented so that by symmetry the period is extremal. If the field lies along or near a (110) axis, the central section cuts four contact regions and in periodically extended reciprocal space the bounding curve is a slightly re-entrant hole orbit whose area is about 60% of the electron orbits we have just considered. There are also non-central extremal (minimal) sections, each enclosing two contact regions, which have distinctly smaller areas still. Oscillations of even longer period should be observed over a wide range of angles, particularly round the (111)

directions, corresponding to orbits enclosing one contact region, but in fact they have never been detected. Nor indeed has an oscillation of any period been found yet in a whisker orientated along (110). It is dangerous to speculate on the meaning of the absence of any particular de Haas–van Alphen oscillation since the factors controlling the amplitude are so little understood, but it is clear that the behaviour around (110) or the long period oscillations, if either can be observed, will prove decisive in the question of whether there is contact with the zone boundary. If the views expressed on p. 204, concerning the non-viability of orbits which pass close to contact regions, should turn out to be correct, it may be impossible to find any oscillations in the critical directions, and the question of contact will remain conjectural. It should not prove impossible to discover whether there are regions of abnormally high scattering probability ; a careful examination

Figure 49. Oscillogram of de Haas–van Alphen oscillations in copper (Shoenberg, unpublished). As the rate of variation of H decreases the oscillations spread out, the period ΔH being virtually constant.

of the magneto-resistance might serve, or, a more elaborate test, a study of the sharpness of the absorption edge in the Kjeldaas ultrasonic experiment when propagation is along a (111) axis. Whether or not these experiments will yield the particular information desired cannot be predicted, but it is certain that there is a strong case for a systematic examination of the properties of copper, by cyclotron resonance, magneto-acoustic effects, high field galvano-magnetic effects, and anything else which might be expected to yield its contribution, however small, to the general picture. The technical problem of preparing good enough samples for these studies has been overcome by several workers, and in other respects copper seems to be almost an ideal metal for a concentrated attack on the elucidation of the electronic structure. It does not seem to be so easy to prepare copper crystals of really low residual resistance as it is with silver or gold, but on the other hand whiskers are more readily produced, and copper surfaces are much more easily electropolished than surfaces of the other noble metals. If the shape of the Fermi surface can be certainly established, systematic measurements on the temperature variation of the de Haas–van Alphen amplitude or on cyclotron resonance (or preferably both) should enable a velocity to be ascribed to every point. If this can be done the way is clear for a serious examination of the properties of copper—

specific heat, transport effects, etc.—to discover the extent to which they are quantitatively intelligible on the basis of a one-particle model.

Between the reasonable expectation of an elucidation of the structure of copper, and the hope for similar success with so much more complex a metal as lead, lies a gulf whose extent can hardly be guessed. It is certainly a measure of the power of the de Haas–van Alphen effect that it can yield any information of value in the latter case; in fact this is to understate what Gold (1958) has achieved. He has shown that the oscillations, of which as many as four might be observed simultaneously, can be analysed into three groups, which he associates with different sheets of the Fermi surface. Not all oscillations are observed over a wide range of angles, and the amplitudes vary considerably and in a complicated manner, all of which is ingeniously incorporated into the interpretation. The form of the Fermi surface is not so much synthesized from the data—it is rather doubtful whether this could be done by a systematic and objective procedure—as adapted by reference to the data from a scheme suggested by the free-electron model. As discussed on p. 186, if a sphere containing 4 electrons per atom is dissected and reconstructed in the various Brillouin zones, a quite remarkable set of surfaces emerges, which with a little smoothing at the corners may be taken to represent what would be expected if the energy discontinuities at the zone boundaries were small. Gold finds that these surfaces account, with comparatively little modification, for most, if not all, of his observations. The first zone is full and therefore contains 2 electrons and contributes no oscillations; the second zone has its maximum energy at the centre, and contains 1.6 electrons, leaving a nearly spherical hole at the centre which gives a rather short period oscillation of isotropic frequency; the third zone is filled only along the edges of the zone, giving in periodically extended reciprocal space the tubular lattice illustrated in figure 13; and lastly the fourth zone has a small number of electrons in each corner which combine in periodically extended space to form a set of cushion-shaped surfaces. There is little comparison with other data possible. The total area may be compared with an estimate based on anomalous skin effect measurements on a polycrystalline sample; even allowing for poor polish on the surface of the latter there seems to be some discrepancy, the total area of Gold's surfaces being distinctly too high. The skin effect measurements certainly need repetition under better conditions before one can conclude that they offer serious objections to Gold's construction. It is also possible to estimate the mean Fermi velocity on each surface from the temperature variation of the amplitude of the oscillations, and hence to calculate the expected value of the specific heat coefficient γ. This comes out about 40% too high, not outside the considerable uncertainty of the estimate.

It is probably not unfair to Gold to characterize his work as an ingenious, not unplausible, but not certainly established delineation of the Fermi surface. It may be that this starting model, the free-electron sphere, is a good approximation to the truth, but if it is not one may doubt whether the data do not contain too many ambiguities of interpretation to make it possible at this stage to proceed further with assurance. In an ideal world one would doubtless settle the monovalent metals first, and then proceed cautiously to investigate the more promising multivalent metals such as beryllium, magnesium, aluminium, etc., in the hope that with increased insight the more difficult metals would yield their secrets. As it is,

however, one can only expect a haphazard attack governed by rather secondary considerations such as the availability of suitable samples. We have tried here to assess the value of various methods, none of which has really been thoroughly tested. It is to be hoped that future work will confirm our general conclusion, that there is now a reasonable expectation of synthesizing from experimental data a valid model of a metal, and will reveal new ways of testing critically the inherent assumptions of the independent-particle model and the theory of transport processes.

APPENDIX

The Standard Metal

We choose as standard a monovalent metal whose atomic volume is 10 cm^3, so that there are 6.0×10^{22} conduction electrons per cm^3. The electrons are assumed to be perfectly free, so that the Fermi surface is a sphere. The magnitudes of the relevant properties are as follows:

Radius of Fermi sphere : $k_0 = 1.21 \times 10^8$ cm^{-1}
Maximal cross-sectional area : $\mathscr{A}_{0,\max} = 4.60 \times 10^{16}$ cm^{-2}
Fermi velocity : $v_0 = 1.40 \times 10^8$ cm sec^{-1}
Cyclotron frequency in a field of H gauss : $\omega_c = 1.76 \times 10^7$ H radians sec^{-1}
Maximal orbit radius in a field of H gauss : $R = 7.96/H$ cm
Electrical conductivity with free path l cm : $\sigma = 1.21 \times 10^{11}$ l $ohm^{-1}\,cm^{-1}$
Electronic specific heat coefficient per unit volume : $\gamma = 632$ erg $deg^{-2}\,cm^{-3}$
Fermi energy : $\epsilon_0 = 8.93 \times 10^{-12}$ ergs
$= 5.57$ ev
Degeneracy temperature : $T_0 = \epsilon_0/k = 64\,700°\,K$

Acknowledgments

During the preparation of this review I have benefited very greatly from discussions with Dr. V. Heine, Dr. J. M. Ziman, Mr. J. E. Aubrey and Mr. A. R. Mackintosh, and from the critical comments of Dr. R. G. Chambers ; to all of whom I express my warmest thanks, while absolving them from any responsibility for its imperfections.

I also thank the authors and publishers of papers from which diagrams have been reproduced, for their permission for publication.

References

Articles distinguished by an asterisk contain extensive references, not all of which are repeated below.

§ 1. Introduction

*Pines, D., 1955, *Solid State Physics* (ed. Seitz and Turnbull) (New York : Academic Press), **1**, 368.

§ 2. The Independent-particle Model

Gold, A. V., 1958, *Phil. Trans. Roy. Soc.* A, **251**, 85.
*Herman, F., 1958, *Rev. Mod. Phys.*, **30**, 102.
Onsager, L., 1952, *Phil. Mag.*, **43**, 1006.
*Reitz, J. R., 1955, *Solid State Physics*, **1**, 2.
*Slater, J. C., 1956, *Hand. d. Phys.* (ed. S. Flügge) (Berlin : Springer), **19**, 1.

§ 3. Transport Phenomena

3.1. *The Method of Trajectories*

Chambers, R. G., 1952 a, *Proc. Phys. Soc.* A, **65**, 458.
Shockley, W., 1950, *Phys. Rev.*, **79**, 191.

3.2. *Spatially Varying Electric Fields*

Pippard, A. B., 1954 a, *Proc. Roy. Soc.* A, **224**, 273.

3.3. *Galvanomagnetic Effects*

Alekseevski, N. E., Brandt, N. B., and Kostina, T. N., 1958, *J. Exper. Theor. Phys.*, **34**, 1339.
Alekseevski, N. E., and Gaidukov, Yu. P., 1958, *J. Exper. Theor. Phys.*, **35**, 554.
Borovik, E. S., 1954, *J. Exper. Theor. Phys.*, **27**, 355; 1955, *Izv. Akad. Nauk, S.S.S.R.*, **19**, 429.
Chambers, R. G., 1956 a, *Proc. Roy. Soc.* A, **238**, 344.
Kapitza, P., 1929, *Proc. Roy. Soc.* A, **123**, 292.
Lifshitz, I. M., 1956, *J. Exper. Theor. Phys.*, **30**, 814.
Lifshitz, I. M., Azbel', M. Ya., and Kaganov, M. I., 1956 a, *J. Exper. Theor. Phys.*, **30**, 220; 1956 b, *Ibid.*, **31**, 63.
Lüthi, B., 1956, *Helv. Phys. Acta*, **29**, 217.
*MacDonald, D. K. C., and Sarginson, K., 1952, *Rep. Progr. Phys.*, **15**, 249 (London: Physical Society).
Okada, T., 1957, *J. Phys. Soc. Japan*, **12**, 1327.
Olsen, J. L., and Rinderer, L., 1954, *Nature, Lond.*, **173**, 682.
Olson, R., and Rodriguez, S., 1957, *Phys. Rev.*, **108**, 1212.
Ziman, J. M., 1958, *Phil. Mag.*, **3**, 1117.

§ 4. Analysis of Special Phenomena

4.1. *Introduction*

*Chambers, R. G., 1956 b, *Canad. J. Phys.*, **34**, 1395.
*Heine, V., and Cohen, M. H., 1958, *Advanc. Phys.* (*Phil. Mag.*, Suppl.), **7**, 395.
*Lax, B., 1958, *Rev. Mod. Phys.*, **30**, 122.
*Pippard, A. B., 1954 b, *Solvay Conf.*, **10**, 123.

4.2. *The de Haas–van Alphen and Related Effects*

Adams, E. N., and Holstein, T. D., 1959, *J. Phys. Chem. Solids*, **10**, 254.
Alers, P. B., 1956, *Phys. Rev.*, **101**, 41.
Alers, P. B., and Webber, R. T., 1953, *Phys. Rev.*, **91**, 1060.
Argyres, P. N., 1958 a, *Phys. Rev.*, **109**, 1115; 1958 b, *J. Phys. Chem. Solids*, **4**, 19.
Argyres, P. M., and Adams, E. N., 1956, *Phys. Rev.*, **104**, 900.
Berlincourt, T. G., 1954, *Phys. Rev.*, **94**, 1172; 1955, *Ibid.*, **99**, 1716.
Berlincourt, T. G., and Steele, M. C., 1954, *Phys. Rev*, **95**, 1421; 1955, *Ibid.*, **98**, 956.
Brailsford, A. D., 1957, *Proc. Phys. Soc.* A, **70**, 275.
Brodie, L. C., 1954, *Phys. Rev.*, **93**, 935.
Chambers, R. G., 1956 b, *Canad. J. Phys.*, **34**, 1395.
Cohen, M. H., and Blount, E. I., 1960, *Phil. Mag.*, in the press.
Connell, R. A., and Marcus, J. A., 1957, *Phys. Rev.*, **107**, 940.
Davydov, B., and Pomeranchuk, I., 1940, *J. Phys. U.S.S.R.*, **2**, 147.
Dhillon, J. S., and Shoenberg, D., 1955, *Phil. Trans. Roy. Soc.* A, **248**, 1.
Fukuroi, T., and Saito, Y., 1957, *Sci. Rep. Res. Insts Tôhoku Univ.* A, **9**, 273.
Gerritsen, A. N., and de Haas, W. J., 1940, *Physica*, **8**, 802.
Gunnersen, E. M., 1956, *Phil. Trans. Roy. Soc.* A, **249**, 299.
Harper, P. G., 1955, *Proc. Phys. Soc.* A, **68**, 874.
Heine, V., 1956 a, *Proc. Phys. Soc.* A, **69**, 505.
Kosevich, A. M., 1957, *J. Exper. Theor. Phys.*, **33**, 735.
Lax, B., Roth, L. M., and Zwerdling, S., 1959, *J. Phys. Chem. Solids*, **8**, 311.

LIFSHITZ, I. M., 1958, *J. Phys. Chem. Solids*, **4**, 11.
LIFSHITZ, I. M., and KOSEVICH, A. M., 1954, *Dokl. Akad. Nauk, S.S.S.R.*, **96**, 963; 1955, *J. Exper. Theor. Phys.*, **29**, 730; 1958, *J. Phys. Chem. Solids*, **4**, 1.
LIFSHITZ, I. M., and POGORELOV, A. V., 1954, *Dokl. Akad. Nauk, S.S.S.R.*, **96**, 1143.
OVERTON, W. C., and BERLINCOURT, T. G., 1955, *Phys. Rev.*, **99**, 1165.
REYNOLDS, J. M., HEMSTREET, H. W., LEINHARDT, T. E., and TRIANTOS, D. D., 1954, *Phys. Rev.*, **96**, 1203.
SCHUBNIKOV, L., and DE HAAS, W. J., 1930, *Commun. Phys. Lab. Univ. Leiden*, **207**, 210.
*SHOENBERG, D., 1952, *Phil. Trans. Roy. Soc.* A, **245**, 1; *1957, *Progress in Low Temperature Physics* (ed. C. J Gorter) (Amsterdam: North-Holland), **2**, 226; 1959, *Nature, Lond.*, **183**, 171.
SLADEK, R. J., 1958, *Phys. Rev.*, **110**, 817.
SOULE, D. E., 1958 a, *Phys. Rev.*, **112**, 698.
STEELE, M. C., and BABISKIN, J., 1955, *Phys. Rev.*, **98**, 359.
VERKIN, B. I., and DMITRENKO, I. M., 1955, *Izv. Akad. Nauk, S.S.S.R.*, **19**, 409.
VERKIN, B. I., DMITRENKO, I. M., and LAZAREV, B. G., 1956, *J. Exper. Theor. Phys.*, **31**, 538.
VERKIN, B. I., DMITRENKO, I. M., and MIKHAILOV, I. F., 1955, *Dokl. Akad. Nauk, S.S.S.R.*, **101**, 233.
ZWERDLING, S., LAX, B., and ROTH, L. M., 1957, *Phys. Rev.*, **108**, 1402.

4.3. *The Anomalous Skin Effect*

CHAMBERS, R. G., 1952 b, *Proc. Roy. Soc.* A, **215**, 481; 1956 a, *Ibid.*, **238**, 344.
FABER, T. E., and PIPPARD, A. B., 1955, *Proc. Roy. Soc.* A, **231**, 336.
FAWCETT, E., 1955, *Proc. Roy. Soc.* A, **232**, 519.
HEINE, V., 1957 a, *Proc. Roy. Soc.* A, **240**, 361.
KAGANOV, M. I., and AZBEL', M. YA., 1955, *Dokl. Akad. Nauk, S.S.S.R.*, **102**, 49.
KLEIN, O., 1944, *Ark. Mat. Astr. Fys.* A, **31**, No. 12.
PIPPARD, A. B., 1954 a, *Proc. Roy. Soc.* A, **224**, 273; *1954 c, *Advances in Electronics and Electron Physics* (ed. L. Marton) (New York: Academic Press), **6**, 1; 1957 a, *Phil. Trans. Roy. Soc.* A, **250**, 325.
REUTER, G. E. H., and SONDHEIMER, E. H., 1948, *Proc. Roy. Soc.* A, **195**, 336.
SMITH, G. E., 1959, *Phys. Rev.*, **115**, 1561.

4.4. *Cyclotron Resonance*

ANDERSON, P. W., 1955, *Phys. Rev.*, **100**, 749.
AUBREY, J. E., and CHAMBERS, R. G., 1957, *J. Phys. Chem. Solids*, **3**, 128.
AZBEL', M. YA., 1958, *J. Phys. Chem. Solids*, **7**, 105.
AZBEL', M. YA., and KANER, E. A., 1956, *J. Exper. Theor. Phys.*, **30**, 811; *1958, *J. Phys. Chem. Solids*, **6**, 113.
BEZUGLYI, P. A., and GALKIN, A. A., 1957, *J. Exper. Theor. Phys.*, **33**, 1076.
CHAMBERS, R. G., 1952 b, *Proc. Roy. Soc.* A, **215**, 481; 1956 c, *Phil. Mag.*, **1**, 459.
DONOVAN, B., and SONDHEIMER, E. H., 1953, *Proc. Phys. Soc.* A, **66**, 823.
DAVYDOV, B., and POMERANCHUK, I., 1940, *J. Phys. U.S.S.R.*, **2**, 147.
FAWCETT, E., 1956, *Phys. Rev.*, **103**, 1582.
GALT, J. K., MERRITT, F. R., YAGER, W. A., and DAIL, H. W., 1959, *Phys. Rev. Lett.*, **2**, 292.
GALT, J. K., YAGER, W. A., and DAIL, H. W., 1956, *Phys. Rev.*, **103**, 1586.
GALT, J. K., YAGER, W. A., MERRITT, F. R., CETLIN, B. B., and BRAILSFORD, A. D., 1960, *Phys. Rev.*, **114**, 1396.
GALT, J. K., YAGER, W. A., MERRITT, F. R., CETLIN, B. B., and DAIL, H. W., 1955, *Phys. Rev.*, **100**, 748.
HEINE, V., 1957 b, *Phys. Rev.*, **107**, 431.
KIP, A. F., LANGENBERG, D. N., ROSENBLUM, B., and WAGONER, G., 1957, *Phys. Rev.*, **108**, 494.
MATTIS, D. C., and DRESSELHAUS, G., 1958, *Phys. Rev.*, **111**, 403.
RODRIGUEZ, S., 1958 a, *Phys. Rev.*, **112**, 1616.

4.5. *Ultrasonic Attenuation and Magneto-acoustic Effects*

BLOUNT, E. I., 1959, *Phys. Rev.*, **114**, 418.
BÖMMEL, H. E., 1954, *Phys. Rev.*, **96**, 220; 1955, *Ibid.*, **100**, 758.
KITTEL, C., 1955, *Acta Metallurg.*, **3**, 295.
KJELDAAS, T., 1959, *Phys. Rev.*, **113**, 1473.
KJELDAAS, T., and HOLSTEIN, T. D., 1959, *Phys. Rev. Lett.*, **2**, 340.
MACKINNON, L., 1955 a, *Phys. Rev.*, **98**, 1181; 1955 b, *Ibid.*, **98**, 1210; 1955 c, *Ibid.*, **100**, 655.
MASON, W. P., 1955, *Phys. Rev.*, **97**, 557; *1958, *Physical Acoustics and the Properties of Solids* (New York: van Nostrand).
MASON, W. P., and BÖMMEL, H. E., 1956, *J. Acoust. Soc. Amer.*, **28**, 930.
MORSE, R. W., 1955, *Phys. Rev.*, **97**, 1716; *1959, *Progress in Cryogenics* (ed. K. Mendelssohn) (London: Heywood), **1**, 219.
MORSE, R. W., and BOHM, H. V., 1957, *Phys. Rev.*, **108**, 1094.
MORSE, R. W., BOHM, H. V., and GAVENDA, J. D., 1958, *Phys. Rev.*, **109**, 1394.
MORSE, R. W., and GAVENDA, J. D., 1959, *Phys. Rev. Lett.*, **2**, 250.
PIPPARD, A. B., 1955, *Phil. Mag.*, **46**, 1104; 1957 b, *Ibid.*, **2**, 1147; 1957 c, *J. Phys. Chem. Solids*, **3**, 175.
RENEKER, D., 1959, *Phys. Rev.*, **115**, 303.
RODRIGUEZ, S., 1958 b, *Phys. Rev.*, **112**, 80.
STEINBERG, M. S., 1958 a, *Phys. Rev.*, **110**, 772; 1958 b, *Ibid.*, **110**, 1467; 1958 c, *Ibid.*, **111**, 425.

4.6. *Size Effects*

AZBEL', M. YA., 1958, *J. Phys. Chem. Solids*, **7**, 105.
BABISKIN, J., and SIEBENMANN, P. G., 1957, *Phys. Rev.*, **107**, 1249.
SONDHEIMER, E. H., 1950, *Phys. Rev.*, **80**, 401; *1952, *Advanc. Phys.* (*Phil. Mag.*, Suppl.), **1**, 1.

§ 5. APPLICATIONS

ABELES, B., and MEIBOOM, S., 1956, *Phys. Rev.*, **101**, 544.
HEINE, V., 1956 b, *Proc. Phys. Soc.* A, **59**, 513.
*MCCLURE, J. W., 1958, *Phys. Rev.*, **112**, 715.
PIPPARD, A. B., 1957 a, *Phil. Trans. Roy. Soc.* A, **250**, 325.
SHOENBERG, D., 1957, *Progress in Low Temperature Physics* (ed. C. J. Gorter) (Amsterdam: North-Holland), **2**, 226; 1959, *Nature, Lond.*, **183**, 171.
*SOULE, D. E., 1958 b, *Phys. Rev.*, **112**, 708.

Note added in Proof.

Since the article was written a very significant advance has been made by Shoenberg (1960), who has continued his examination of the de Haas–van Alphen effect in copper, silver and gold to the point where the general forms of the Fermi surface have become clear. The variety of oscillatory periods and their angular dependence leave little doubt that there is contact with the zone boundary around the (111) position in all three, of the type shown in figure 48. The area of contact is greatest in copper and least in silver. It appears that the contact region in copper is about 12% greater in diameter than was suggested by the anomalous skin effect, but this is probably not a serious discrepancy since the latter is not sensitive to small variations in this region. There are other discrepancies in detail which are still unresolved, and will remain so until a more extensive body of de Haas–van Alphen data has been collected and a thorough analysis made of all alternative interpretations. Recent measurements of the anomalous skin effect in silver by

V. M. Morton reveal a variation of resistance which is similar to that of copper but over a smaller range; this agrees with Shoenberg's suggestion of a smaller contact area, but it must be admitted that in the absence of his evidence for contact one would have been reluctant to infer it from the skin effect data. Nevertheless, however unwilling one may be to accept the idea that the same independent-particle model can be made to fit all the facts exactly, it is surely very unlikely that one set of facts will require contact and another set not. Clearly a good deal of examination of alternative models will be needed here also, and it will be particularly important to try to estimate the range of shapes which are consistent with the skin effect data.

Cyclotron resonance has been observed in copper by Langenberg and Moore (1959); the results are not sufficiently detailed yet to allow a clear interpretation, but it is interesting to note that the cyclotron frequency shows singularities at certain orientations of the magnetic field which may be associated with a change in character of the dominant orbits according as the sections of the Fermi surface enclose electron or hole orbits. This behaviour indicates that studies of cyclotron resonance may yield topological information about the connectivity of the Fermi surface, quite apart from the quantitative information about the Fermi velocity which is discussed in the text. The qualitative problem, of finding what sort of surface the Fermi surface is, has in fact been rather neglected in the article, and this recent work serves as a reminder of the fundamental importance of determining whether and where zone boundary contacts occur.

But perhaps the most serious error in the article is one both of fact and of emphasis. It has been pointed out by Lifshitz and Peschanskii (1958) that open orbits are not confined to magnetic field orientations which lie in planes normal to the high symmetry axes discussed on p. 203. If **H** is oriented in such a way (say rather close to (100) or (110) or (111)) that some plane sections normal to **H** define electron orbits and some hole orbits, then there must be an intermediate set of sections which define open orbits. These latter are not regularly periodic trajectories of the type shown in figure 14, but are formed of a succession of different partial revolutions of the Fermi surface at different levels, such as will allow the general trend of the trajectory in reciprocal space to be linear but not parallel to a symmetry axis. This analysis crystallizes the meaning of Chambers' (1956 b) 'quasi-ergodic' orbits, and shows them often to be genuinely open. For a Fermi surface such as that of copper open orbits are by no means a rarity, though they are confined to orientations of **H** which lie within cones centred on the axes of high symmetry. It is therefore possible to make a start at interpreting Gaidukov's (1959) most valuable observations on the transverse magneto-resistance of the noble metals, which fails to saturate when **H** lies within just such cones as these. Indeed Priestley (1960) has shown that the details of the observations of saturation and non-saturation in gold are almost quantitatively explained on the basis of Shoenberg's estimate of the size of the contact regions in the (111) directions. As a result of this Russian work it appears for the first time as if a complete account of the high-field magneto-resistance may be given along quasi-classical lines (or of course the more sophisticated approaches mentioned in the text). This in itself is a conclusion of great importance, for, if justified, it removes what appeared to be a serious impediment to the understanding of transport phenomena in metals. In addition, the essential simplicity of the topological argument of

Lifshitz and Peschanskii opens the door for the systematic use of magneto-resistance for determining the connectivity of the Fermi surface—a most attractive development in view of the elementary character of the experimental techniques involved.

Further experimental work by Gavenda and Morse (1959) on magneto-acoustic oscillations in copper has revealed a long-period oscillation which may be interpreted as an orbit around the contact region on the zone-boundary, the estimated size agreeing well with Shoenberg's estimate. In view of the discussion on p. 203 it should not be a surprise to find that a detailed analysis of the effect in metals of arbitrary Fermi surface shows how the attenuation and its oscillations, particularly for longitudinal waves, are dominated by any regions which are especially sensitive to deformation, as one might expect the contact regions to be. It seems likely, in fact, that the apparently anomalous periodicity of the oscillations in figure 43 are to be explained in this way; they do not arise from a central orbit, but from a rather squarer orbit (or a rosette-like hole orbit of the sort found by Shoenberg) which passes near four contact regions. The dimensions of such an orbit are consistent with the observed periodicity.

Our conclusion at the moment, then, is that there is every reason for optimism. Unless something unexpected happens the electronic structure of the noble metals should be fairly exactly known within a very few years, and the hoped-for test of the basic assumptions should then be possible. Moreover, it seems that there is a good chance that the independent-particle model will prove a close approximation to the truth.

Gaidukov, Yu. P., 1959, *J. Exper. Theor. Phys.*, **37**, 1281.
Gavenda, J. D., and Morse, R. W., 1959, *Bull. Amer. Phys. Soc.*, **4**, 463.
Langenberg, D. N., and Moore, T. W., 1959, *Phys. Rev. Lett.*, **3**, 328.
Lifshitz, I. M., and Peschanskii, V. G., 1958, *J. Exper. Theor. Phys.*, **35**, 1251.
Priestley, M. G., 1960, *Phil. Mag.*, in the press.
Shoenberg, D., 1960, *Phil. Mag.*, in the press.

NEW DEVELOPMENTS IN INTERFERENCE SPECTROSCOPY

By P. JACQUINOT

University of Paris and Laboratoire Aimé Cotton,
Centre Nationale de la Recherche Scientifique, Bellevue, France

CONTENTS

PART I. INTRODUCTION PAGE

§ 1. Introduction 268

PART II. SPECTROSCOPY BY MEANS OF THE FABRY–PEROT APPARATUS

§ 2. General remarks 271
2.1. Spectrographic method 272
2.2. Application to spectrometry 272
§ 3. Instrumental function 272
3.1. Effect of the decrease in the amplitude of the successive transmitted rays 272
3.2. Effect of surface imperfections 273
3.3. Effect of the use of a finite range of incidences 273
3.4. Actual case 274
§ 4. Resolving power and finesse 274
§ 5. The F.P. monochromator and the F.P. spectrometer 275
5.1. The luminosity of the F.P. spectrometer 276
5.2. Exploration of the spectrum 278
§ 6. The F.P. spectrometer with a single bandpass 280
§ 7. Technical considerations 282
7.1. Coatings 282
7.2. Materials 285
7.3. Mountings 285
7.4. General arrangement 285
§ 8. The spherical Fabry–Perot 286
§ 9. The photographic use of the F.P. 287

PART III. METHODS MAKING USE OF THE AMPLITUDE OF INTERFEROMETRIC MODULATION

§ 10. Connes' method 288
10.1. Michelson interferometer 289
10.2. SISAM : instrumental function and theoretical resolving power 290
10.3. Permissible solid angle 292
10.4. Effect of unmodulated light 293
10.5. Practical considerations on the use of SISAM 294
10.6. First results 294
§ 11. Guilino and Maier-Leibnitz' method 294

PART IV. METHODS MAKING USE OF THE FREQUENCY OF INTERFEROMETRIC MODULATION. SPECTROSCOPY BY FOURIER TRANSFORMATION

§ 12. General principles of the method 295
§ 13. Resolving power 296
§ 14. Admissible solid angle 297
§ 15. Fundamental advantage of the method 298

PAGE

§ 16. Transformation of the interferogram into the spectrum 299
16.1. Direct use of electric filters 299
16.2. Instantaneous analogue computer of Strong and Vanasse 299
16.3. Numerical transformation 300
16.4. Application of specialized computers 302
§ 17. Experimental results 303
§ 18. Michelson's method of fringe visibility 305

PART V. MISCELLANEOUS

§ 19. Michelson's echelon and Harrison's echelle 307
§ 20. Lummer plate 309
§ 21. The basic reasons for the superiority of certain devices 309
Acknowledgments 310
References 310

Abstract. After an introduction (Part I) dealing with the different possibilities of classifying the methods of interference spectroscopy, Part II deals with the more recent developments in Fabry–Perot spectroscopy. A complete theory, including the role of surface imperfections and of field diaphragm, is given, and the recent improvements in dielectric coating are reviewed. A complete outline of the photoelectric use of the Fabry–Perot (Fabry–Perot ' spectrometer ') is given, including: methods of scanning for low and high resolutions, best compromise between luminosity and resolution, calculation of the luminosity and extensions of the Fabry–Perot method to studies of absorption spectra. The properties of a new ' spherical Fabry–Perot ' spectrometer are described. Some recent results obtained with photographic methods are briefly discussed.

Part III deals with a new type of spectrometer which selects the different wavelengths by the *amplitude* of the modulation given to them by a linear variation of optical path. Such an instrument using gratings, for instance, is capable of yielding the same resolution as would be obtained if one of the gratings was used in a conventional fashion with very narrow slits, but it provides a much higher luminosity. The spectrum is directly recorded in the usual form.

Part IV deals with methods of selecting the different wavelengths by the *frequency* of their interferometric modulation: the spectrum is not directly given in the usual form, but in the form of interferogram which is the Fourier transform of the spectrum. The main advantage of these methods is that, in common with the photographic method, the time required for recording the interferogram does not depend on the width of the spectrum studied. The resolving power and the luminosity of the process are studied. Some methods of application are described, as well as the different methods of reconstitution of the spectrum from the interferogram. Some recent applications of Michelson's method of ' visibility of fringes ' are also reviewed.

PART I. INTRODUCTION

THIS report is concerned with the development of interference spectroscopy during the last ten or fifteen years. After a long period of stability, the techniques of interference spectroscopy have recently recovered great interest, and a part of the considerable work on this question was presented at an International Colloquium organized by the Centre National de la Recherche Scientifique held at Bellevue, France, in September 1957. Before this colloquium, in 1951, an excellent article was published in *Reports on Progress in Physics* entitled ' New Techniques in Optical Interferometry' by Kuhn. Since that date new progress has been made; furthermore the report of Kuhn was concerned with

all the techniques of interferometry while this report will deal particularly with interference spectroscopy.

It is relatively difficult to give a good definition of interference spectroscopy. In reality, all the techniques of spectroscopy involve the phenomena of interference. In the prism the dispersion of light is due to the fact that different wavelengths are subject to different optical retardations in the prism material, and the resulting interference is constructive only in the directions given by the elementary laws of the prism; this is a phenomenon of interference of zero order. In gratings one obtains dispersion of light in directions which correspond to orders of interference different from zero, but equal only to several units, or exceptionally equal to several hundreds as in the echelon grating. On the other hand, in the instruments such as the Lummer plate or the Michelson echelon, the Fabry–Perot interferometer or yet other methods based on the Michelson interferometer, there is also an interference phenomenon which permits the separation of different radiations, but here the order of interference is much higher, and may attain several tens or hundreds of thousands. We thus reserve the name of interference spectroscopy for techniques which use high orders of interference; this leads us, in a rather arbitrary manner it must be admitted, to exclude gratings from our survey.

Several methods of classification of interference phenomena exist:

One can, for example, divide interference phenomena into two classes according to the number of interfering rays, two or several. For example gratings or the Fabry–Perot étalon are cases of multiple-beam interferometry; the Michelson interferometer, which can also be employed in modern spectroscopic methods, is one of the simplest examples of two-beam interferometry.

Interference phenomena can also be divided into two classes according as the division of the incident energy into coherent beams, which are then made to interfere, is made by a spatial division (e.g. the Young slits experiment, gratings) or by a splitting of the luminance of each ray by a device such as a semi-transparent plate (e.g. Lummer plate, Fabry–Perot étalon, Michelson interferometer). We shall see in what follows that it is the second class of phenomenon, that is to say the division of luminance, which has the greatest possibilities in interference spectroscopy.

It will be evident at the end of this report that one can also divide the techniques of interference *spectroscopy* into two classes, according to whether the separation of various radiations is made by means of selection in space (gratings, Fabry–Perot spectrograph, Lummer plate, etc.), or whether the different radiations are distinguished from one another by one of the qualities of the modulation which is imposed on the light beam by a continuous variation of the optical retardation.

Actually the adoption for this report of either of these methods of classification would lead to rather arbitrary divisions. It appeared more simple to choose the following plan.

(1) The Fabry–Perot instrument.
(2) Methods using the amplitude of interferometric modulation.
(3) Methods using the frequency of interferometric modulation.
(4) Miscellaneous.
(5) General remarks.

In this report we shall consider a certain number of characteristics common to all the devices and of which the principal is the instrumental line shape function (hereafter referred to as ' instrumental function ').

The properties of resolution of a spectral device are completely determined by the representation which it gives of a perfectly monochromatic spectral line. If the total luminance

$$\int_0^\infty B(\sigma)\,d\sigma, \quad (\sigma = 1/\lambda = \text{wave number})$$

is localized in a frequency interval of zero width ($B(\sigma) = \delta(\sigma - \sigma_0)$, δ being the Dirac function), the representation of this line by the apparatus is spread out over a small domain of frequency according to a function $B'(\sigma) = W(\sigma - \sigma_0)$ which is by definition the instrumental function. If now the apparatus is illuminated by radiation of spectral density $B(\sigma)$ (the source function), it will yield a representation $B'(\sigma)$ which differs from the real spectrum $B(\sigma)$:

$$B'(\sigma) = \int_0^\infty W(\sigma - \sigma_0)\,B(\sigma_0)\,d\sigma_0.$$

$B'(\sigma)$ is the convolution of the source function by the instrumental function. This may be written $B'(\sigma) = W(\sigma) \star B(\sigma)$.

$B'(\sigma)$ resembles more and more closely $B(\sigma)$ as the ' width ' w of W becomes smaller in comparison with the width of variations in $B(\sigma)$. Furthermore two monochromatic lines are resolved when their wave number separation $\sigma_2 - \sigma_1$ is greater than a certain value $\delta\sigma$, which is of the order of the width w of W. $\delta\sigma$ is the *effective* limit of resolution. The relation between $\delta\sigma$ and w depends evidently on the criteria adopted for the resolution, the form of the function W and the definition of its ' width ' w. In order not to complicate the discussion by unimportant details we will adopt $\delta\sigma = w$, and we will choose for the definition of w the one which leads to the simplest results in each particular case ; this will be, for example, the width at half height for a Gaussian curve, or the distance between the maximum and the first zero for a diffraction curve. The results given by this method are found more simply and differ very little from those which are obtained by more rigorous definitions ; even these definitions are in a certain sense arbitrary.

The resolving power, or *resolvance*, is always defined by $\mathscr{R} = \sigma/\delta\sigma = \lambda/\delta\lambda$. It must be clearly understood that this is the *effective* resolvance, that is to say the resolvance that is obtained taking into account all the characteristics of the device which contribute to the width of W. If certain of these characteristics are idealized, for example the use of infinitely narrow slits and the existence of perfect optical surfaces, $\mathscr{R}$ is only limited by fundamental phenomena and it takes a value $\mathscr{R}_0 = \sigma/\delta\sigma_0$ called the theoretical resolvance. We shall see that the theoretical resolvance is expressed by the general relation $\mathscr{R}_0 = \Delta_M/\lambda$, Δ_M being the maximum optical retardation between the interfering beams. This expression, correct for most interference apparatus, is a particular case of the more general expression $\mathscr{R}_0 = d\Delta_M/d\lambda$ applicable, for example, to the prism.

Another important characteristic is the luminosity of the device. The luminosity is proportional to the product U of the area of the cross section of the outgoing beam and the solid angle Ω associated with the resolvance $\mathscr{R}$ (U in current French usage is the ' étendue ' ; in English, the light-gathering power).

In nearly all devices, if one tries to change $\mathscr{R}$ one finds that the product $U\mathscr{R}$ remains constant and is equal to the product of the transverse dimensions of the optical system and a constant, which depends only on the method used and can well be greater for certain methods than for others.

PART II. SPECTROSCOPY BY MEANS OF THE FABRY-PEROT APPARATUS

§ 2. General Remarks

This apparatus, invented in 1896, has played a considerable role in high resolution spectroscopy, and the majority of investigations of hyperfine structure have been made by means of the Fabry–Perot étalon and the photographic method. For a long time the apparatus was very little developed and its use was limited to this kind of investigation. It was not realized until recently that it could be used with success outside the photographic region, and over a large range of resolution; considerable progress, both theoretical and practical, has been made.

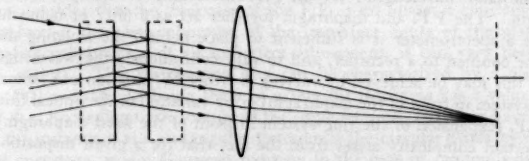


Figure 1. The Fabry–Perot interferometer.

We will thus emphasize primarily the properties of the apparatus which have only recently been brought to light, and the new applications.

The elementary theory of the apparatus is well known but it is somewhat insufficient to explain completely the new uses. Before explaining the recent progress we will recall very briefly this theory with some additional refinements.

The Fabry–Perot element (sometimes called étalon, and which we will designate henceforth simply by F.P.) consists of two transparent plates of which the faces are, in principle, perfectly plane and parallel. Each one of these faces is covered by a reflecting coating of reflection coefficient R, transmission coefficient T, and coefficient of absorption $A = 1 - R - T$. An incident ray is divided by multiple transmissions and reflections into an infinity of rays reflected and transmitted parallel to each other. The transmitted rays, the only ones used in general, recombine at infinity and interfere. At the focus of a lens (figure 1) one can thus see interference fringes.

Between two successive transmitted rays there exists an optical retardation Δ. If the ratio Δ/λ is an integer the transmitted rays interfere constructively and the incident energy is totally transmitted. If this condition is not exactly fulfilled the multiple rays interfere destructively, the constructive interference of a great number of vibrations being very critical, and the incident energy is reflected back to the source. For a given value of Δ the F.P. acts as a wavelength filter; the variation of Δ with the thickness of the F.P., or with the angle of incidence of the beam, thus permits the spectroscopic application of the F.P.

2.1. *Spectrographic Method*

The optical retardation has the value $\Delta = 2nt\cos i$, nt being the optical thickness of the F.P. and i the angle of incidence of the beam. Since Δ depends on i, the different radiations of the spectrum are transmitted in different directions and a photographic plate placed in the focal plane receives at each point the corresponding information; the interpretation of the photograph obtained is sometimes rather difficult. Each radiation gives, in contrast to the case of the prism, not a line image, but a system of rings centred on the normal to the F.P. (the loci of points for which $\cos i = K\lambda/2nt$). We shall see later how one obtains an unambiguous knowledge of the spectrum.

2.2. *Application to Spectrometry*

The application to spectrometry consists, generally, of the replacement of the photographic plate by a diaphragm which isolates a region of points which receive rays of the same wavelength: for the F.P. this could, for example, be an annular diaphragm. The F.P. and diaphragm together act as a filter or monochromator. To make a spectrometer it is sufficient to place behind the isolating diaphragm a detector coupled to a recorder, and to vary continuously the wavelength transmitted; this may be achieved by varying Δ either by variation of i (translation of the ring system in front of the diaphragm) or by variation of the optical thickness nt of the F.P. (expansion of the ring system in front of the fixed diaphragm). Here also, a certain complexity arises from the fact that for a given disposition of the diaphragm and the ring system a whole series of wavelengths are accepted, $\lambda_K = \Delta/K$, K being any integer.

§ 3. Instrumental Function

In the classical theory of the F.P. it is assumed that the plates are perfectly plane and parallel, and that the width of the diaphragm, or the grain size of the plates in the focal plane are negligible. Of course the conditions are never fulfilled and it is necessary to have a more complete theory. Chabbal (1953) has shown that the instrumental function is not the well-known Airy function, but the convolution of three functions which we shall now define.

3.1. *Effect of the Decrease in the Amplitude of the Successive Transmitted Rays*

Let us now assume that the plates are perfectly plane and parallel and that the above-mentioned diaphragm has no width: then all the interfering rays have the well-defined retardation $\Delta = 2nt\cos i$. If the reflecting power were equal to unity, all these rays would have the same amplitude, and, their number being infinite, the interference should be destructive for all the wave numbers except for $\sigma_K = K/\Delta$. Under these conditions the instrumental function would be presented by a series of infinitely sharp peaks (Dirac peaks)

$$W(\sigma) = \sum_{0}^{\infty} \delta(\sigma - \sigma_K)$$

separated by the interval $\Delta\sigma = 1/\Delta = 1/2nt\cos i$. With a reflecting power $R<1$ the instrumental function is the Airy function

$$A(\sigma) = \left(\frac{T}{1-R}\right)^2 \frac{1}{1+(4R\sin^2\pi\sigma\Delta)/(1-R)^2}.$$

Each peak has now a width $a = \Delta\sigma/N_R$, $N_R = \pi\sqrt{R}/(1-R)$ being a coefficient which does not depend on the thickness; its name 'reflecting finesse' means that it depends only on the reflecting power. The transmission has the maximum value

$$\tau_A = \left(\frac{T}{1-R}\right)^2 = \left(1-\frac{A}{1-R}\right)^2;$$

it differs from unity only because the coatings have an absorption coefficient $A = 1-T-R$.

3.2. *Effect of Surface Imperfections*

Let us again assume that $R = 1$ and that all the rays have the same incidence, but let us consider that the thickness of the F.P. is no longer a constant over the whole surface of the plate: because of the imperfect figure of the plates, or of an imperfect polishing or of an imperfect adjustment, it can vary between the values t_1 and t_2. This distribution of t can be characterized by a distribution function $D(t) = dS/dt$, dS being the area where the thickness has a value between t and $t+dt$. According to the nature of the imperfections, $D(t)$ may have different forms: for instance it is a rectangular function if the circular plates have a slight spherical curvature, it is a Gaussian function in the case of randomly distributed micro defects; in any case the shape of $D(t)$ can be deduced from experimental measurements on the plates. It is now evident that the Dirac peaks of the ideal instrumental function are replaced by peaks of the form $D(\sigma-\sigma_0)$, a function deduced from $D(t)$ by the relation $\sigma = 1/2nt\cos i$. The width d of the peak of $D(\sigma-\sigma_0)$ is related to the width δt of $D(t)$ by $d/\sigma = \delta t/t$ or $d = \Delta\sigma.2\delta t/\lambda$. The ratio of d to $\Delta\sigma$ is also a coefficient of finesse of the same nature as N_R, $N_D = d/\Delta\sigma = 2\delta t/\lambda$; it is probably the most severe limitation and for this reason has been called 'limiting finesse' by Chabbal. If δt is expressed in terms of λ, ($\delta t = \lambda/m$), $N_D = m/2$: for example if the plates are flat, to within $\lambda/50$, the finesse cannot exceed 25.

3.3. *Effect of the Use of a Finite Range of Incidences*

Similarly, let us assume that $R = 1$, ($a = 0$, $N_R = \infty$) and that the plates are perfect ($d = 0$, $N_D = \infty$); but let us take into account the finite width of the focal diaphragm, or the finite size of the domain of integration of the photographic plate. Thus different angles of incidence contribute to the result: this can be characterized by a distribution function $F(\cos i) = d\Omega/d(\cos i)$, the contribution of the rays with incidences $(i, i+di)$ being proportional to their solid angle $d\Omega$. The shape of $F(\cos i)$ depends on the nature of the diaphragm (or plate): in the particular case of an annular diaphragm it is a rectangular function of width $\delta(\cos i) = \cos i_1 - \cos i_2$. The instrumental function is now composed of peaks, with the same interval $\Delta\sigma$, each peak having the form $F(\sigma-\sigma_0)$, a function deduced from $F(\cos i)$ by the relation $\sigma = 1/2nt\cos i$. The peak width f is deduced from

the width of $F(\cos i)$ by the same relation and is $f = \sigma\delta(\cos i)/\cos i \simeq \sigma\delta(\cos i)$. In the important case of the annular diaphragm, or of the circular hole generally used in the F.P. spectrometer, $d\Omega = 2\pi d(\cos i)$. The solid angle Ω associated with the diaphragm is thus $\Omega = 2\pi f/\sigma = 2\pi/\mathscr{R}_F$, $\mathscr{R}_F$ being the resolving power that would be obtained with a perfectly plane and perfectly reflecting F.P. (using this solid angle). This leads to the important relation $\Omega\mathscr{R}_F = 2\pi$.

3.4. *Actual Case*

In the actual case all the three effects may be of the same order of magnitude. With only two effects, say A and D, one would obtain an instrumental function

$$W'(\sigma) = \int A(\sigma')\, D(\sigma - \sigma')\, d\sigma' = A(\sigma) \star D(\sigma).$$

The addition of the third effect leads to

$$W(\sigma) = \int W'(\sigma')\, F(\sigma - \sigma')\, d\sigma' = A(\sigma) \star D(\sigma) \star F(\sigma) :$$

thus the instrumental function is given by the convolution of three elementary functions A, D and F, which play symmetrical roles.

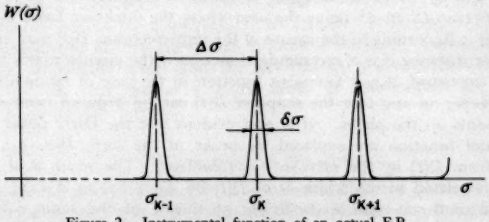


Figure 2. Instrumental function of an actual F.P.

§ 4. Resolving Power and Finesse

The limit of resolution is the width w of the instrumental function W. The ratio σ/w of the wave number to w is the resolvance $\mathscr{R} = \sigma/\delta\sigma = \sigma/w$ while the ratio of the free spectral range $\Delta\sigma$ to w is the *effective finesse* $N = \Delta\sigma/w$. Moreover $\mathscr{R} = (\sigma/\Delta\sigma)(\Delta\sigma/w) = KN = N.2nt/\lambda$. Thus N is the resolvance per unit interference order. It is also the maximum number of peaks of W which can be resolved in the interval of two adjacent interference orders, and thus the maximum number of spectral elements which can be observed : for this reason it plays a very important role in the use of the F.P.

According to the formula $\mathscr{R} = KN$ the resolvance can be increased by increasing either K or N. Of course it would be better to increase N than K, but while the thickness (and hence K) can be indefinitely increased, the same does not hold for N.

In effect $W(\sigma)$ is the convolution of $A(\sigma), D(\sigma)$ and $F(\sigma)$ and its width w is a function of the widths a, d, f which may be symbolically written $w = a \oplus d \oplus f$. Such functions have been studied by many authors among whom are Born (1933),

Minkowski and Bruck (1935), Chabbal (1953). The exact relation between w and a, d, f, depends on the particular shape of the functions D and F, but the general conclusions are :

(i) w is greater than any of the three values a, d, f.

(ii) w is smaller than the sum $a+d+f$.

(iii) If one of the values is zero (for instance $f = 0$), and the two others are equal (for instance $a = d$), then $w \simeq 1{\cdot}4a$.

(iv) If $a = d = f$, then $w \simeq 1{\cdot}7a$.

Consequently, the finesse $N = \Delta\sigma/w$ is smaller than any of the three values $N_R = \Delta\sigma/a, N_D = \Delta\sigma/d, N_F = \Delta\sigma/f$: its value is $N \simeq 0{\cdot}7N_R$ if $N_R = N_D$ and $f = 0$ and $N \simeq 0{\cdot}6N_R$ if $N_R = N_D = N_F$. As an example figure 3 shows the variation of N/N_D as a function of N_R/N_D when $f = 0$ in the case where $D(\sigma)$ is a Gaussian function or a rectangular function.

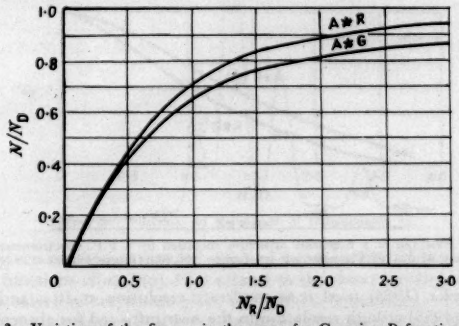


Figure 3. Variation of the finesse, in the case of a Gaussian D function ($A \star G$) and in the case of a rectangular function ($A \star R$).

$N_F = \Delta\sigma/f$ can be increased indefinitely ; N_R also can reach very high values, up to several hundreds for example. But, as has been stated above, N_D is limited to $m/2$ if the plates are worked to within λ/m. Thus the effective finesse N is limited to the value of N_D which, in the visible region† can hardly exceed 50. This limitation of N to rather low values is actually the greatest disadvantage of the F.P. The most important improvement which can now be expected is in the figuring of the plates ; it is not impossible to hope that technical progress will permit a noticeable increase in the values of N. Fortunately there exist methods of increasing artificially the value of N by placing two F.P. in series, as will be discussed in § 6.

§ 5. The F.P. Monochromator and the F.P. Spectrometer

The monochromator and spectrometer essentially consist of a pair of Fabry–Perot plates, and a diaphragm in the focal plane of a lens placed behind the F.P. :

† In the microwave region, very high performances have been reached by Culshaw, Richardson and Kerns (1959).

usually the diaphragm is circular. Furthermore, a premonochromator is very often placed between the source and the F.P.

A simplified form of the F.P. monochromator is the interference filter commonly used for very low resolution. In that case the layer separating the reflecting planes is solid, generally made of evaporated cryolite or zinc sulphide ; however, in some cases the conventional structure of the F.P. étalon has been used, with very thin spacers.

The F.P. spectrometer was first introduced by Jacquinot and Dufour (1948). Since that date it has been frequently used and improved : Blaise (1953 etc.), Chantrel (1958, 1959), Chabbal (1954), Kaliteevsky and Tchaika (1956), Kuhn (1958), Lennuier and Cojan (1952), Brody, Tomkins and Fred (1957) amongst others have used it for hyperfine structure studies, in the visible (figure 4), ultra-violet and infra-red regions. Dupeyrat (1957) has obtained results on the Raman

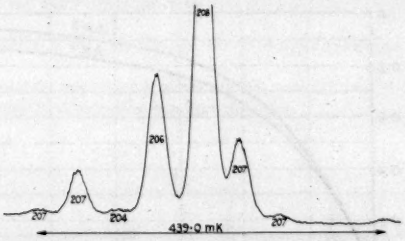


Figure 4. Example of a hyperfine structure recorded by a F.P. spectrometer. Line 4168 Å of Pb, effective resolvance 950,000 (Blaise 1958).

effect. Greenler (1958) used it at moderate resolution at 10 μ and Jaffe, Rank and Wiggins (1955) at high resolution in the near infra-red for absorption spectroscopy. Geake and Wilcock (1956), Ring and Woolf (1958) have studied its uses in astronomy, and Hirschberg (1958) in atmospheric physics.

The success of the F.P. spectrometer is due to several factors. It has over the F.P. spectrograph the well-known advantages of the methods using physical detectors, namely the extension to the infra-red region. Moreover the F.P. spectrometer turned out to be much more luminous than conventional prism or grating spectrometers with the same resolving power, regardless of the value of this resolving power. We now examine two points : the luminosity and the methods for scanning the spectrum.

5.1. *The Luminosity of the F.P. Spectrometer*

The luminosity of a spectrometer is defined as the ratio $L = \phi/B$ of the flux falling on the detector to the luminance of the source. It is $L = S\Omega\tau$ where S is the surface area of the plates, Ω the solid angle limited by the focal diaphragm and τ a transmission factor which is, itself, the product of three factors

$$\tau = \tau_A \tau_E \tau_F.$$

τ_A, already defined in § 3.1, depends only on the coatings.

τ_E is a factor due to the convolution $A \star D$. It should be 1 if the surfaces were perfect, and varies as a function of $d/a = N_R/N_D$ according to figure 5 : its value is about 0·75 when $N_R = N_D$ and decreases rapidly when $N_R > N_D$.

τ_F is a factor due to the convolution $(A \star D) \star F$. It should be 1 with a diaphragm of zero width and decreases as the ratio $f/(a \oplus d)$ increases : its value is about 0·8 when $f = a \oplus d$.

Jacquinot and Dufour for the case of a perfect F.P., and Chabbal for the general case, have examined the conditions to be fulfilled in order to have the maximum luminosity and have given the value of this maximum luminosity as a function of the resolvance.

The resolvance and the luminosity can be changed by changing one or several of the parameters which characterize the spectrometer : thickness, reflecting power, solid angle, area utilized on the plates (on which may depend the value

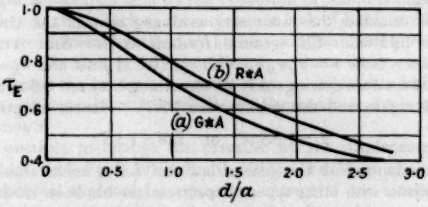


Figure 5. Variation of the factor of transmission τ_E.

of N_D). A given resolvance may be obtained by an infinity of sets of the widths a, d, f of the functions $A(\sigma), D(\sigma), F(\sigma)$ related to the above mentioned-parameters. But there is only one set which gives the highest luminosity. This set would correspond to the condition $a = d = f$ if the three functions A, F and D had the same shape ; in fact the difference in shape of the three functions introduces coefficients not differing much from unity and the equality of the three widths may be adopted practically as the condition for optimum luminosity.

In this case the effective resolvance $\mathscr{R}$ is about $0{\cdot}6\mathscr{R}_F = 0{\cdot}6\mathscr{R}_D = 0{\cdot}6\mathscr{R}_A$, $\mathscr{R}_F$, $\mathscr{R}_D$, $\mathscr{R}_A$ being the resolvance that would be obtained if two of the three widths were zero. Thus a F.P. interferometer, in contrast to a conventional prism or grating spectrometer, is correctly used only if its effective resolvance is of the same order as the theoretical resolvance : the same condition holds for the other interference spectrographs or spectrometers.

The practical conclusions are :

(*a*) The coating must be suitably chosen so as to have $a \simeq d$, or, in other words $N_R = N_D$; thus the resulting finesse N is about $0{\cdot}7N_D$, and $\tau_E = 0{\cdot}8$. If the coating is not well adapted to the plates, for example a coating with too high a reflectivity giving $N_R = 3N_D$, the resulting finesse is only very slightly increased ($N \simeq 0{\cdot}9N_D$) and the transmission factor drops to $\tau_E \simeq 0{\cdot}3$. These conclusions have been used by Wark and Stone (1955).

(*b*) The focal diaphragm must be chosen so as to have $f \simeq a$: it must be changed if a different resolvance is required (either by change of the thickness or by change

of the coating). The mode of use of a F.P. differs noticeably from that of a conventional spectrometer, where the change of resolvance is obtained only by changing the slit widths.

The full luminosity of the F.P. spectrometer is obtained only if all the rays having the same incidence are used, that is to say if the focal diaphragm is annular. This annular diaphragm is almost always reduced to a circular hole, which is an annular diaphragm of zero internal radius. From the relation $\Omega = 2\pi/\mathscr{R}_F$ it can be seen that the angular diameter of the hole is $\alpha = 2\sqrt{2}/\sqrt{\mathscr{R}_F} = 2{\cdot}2/\sqrt{\mathscr{R}}$. Of course it would be possible to use several concentric annular diaphragms as was proposed by Jacquinot and Dufour (1948) but this possibility does not seem ever to have been used. An incidental advantage of using a circular diaphragm is that the rays are nearly normal to the F.P. ; thus the limitation of the number of rays caused by the obliquity in case of high resolution (Van der Sluis and MacNally 1956) is avoided.

Let us now calculate the luminosity assuming that all the above mentioned conditions are fulfilled. The general formula is $L = S\Omega\tau$ with $\Omega = 2\pi/\mathscr{R}_F$ and $\mathscr{R}_F = \mathscr{R}/0{\cdot}6$. Now $\tau = \tau_A \tau_E \tau_F$ with $\tau_F \simeq 0{\cdot}8$ if $f \simeq a$ and $\tau_E = 0{\cdot}75$ if $d \simeq a$. τ_A with good modern coatings may be something between 0·5 and 0·9 according to the spectral region and the quality (see § 7.1). Hence, adopting an average value $\tau_A = 0{\cdot}7$

$$L \simeq 1{\cdot}5S/\mathscr{R} \quad \text{or} \quad L\mathscr{R} \simeq 1{\cdot}5S$$

which is in accordance with the general law that $L\mathscr{R}$ is an invariant.

The comparison with other types of spectrometer has been made by Jacquinot (1954 b) ; let us consider only the comparison with a grating spectrometer, which is more luminous than a prism spectrometer. It is easily found that the luminosity of the grating is $L' = K\tau S'\beta(\sin i_1 + \sin i_2)/\mathscr{R}$. In this formula i_1 and i_2 are the angles of incidence and emergence, β the angular height of the slit, S' the area of the grating, τ the transmission factor which may vary between 0·25 and 0·75, and K a factor varying from 0 if $\mathscr{R} \simeq \mathscr{R}_0$ to 1 if $\mathscr{R} \ll \mathscr{R}_0$. For comparison, let us assume that $\beta = 1/50$, $K = 1$, $\tau = 0{\cdot}4$, then $L/L' = 200S/S'$ for the same $\mathscr{R}$. Thus for the same area, the gain in luminosity would be very great : one could also say that the same luminosity would be obtained with a F.P. of area 200 times smaller than a grating.

In this discussion it is assumed that the F.P. is not associated with any pre-monochromator. If such a monochromator is used, the gain is decreased, not only because the monochromator has a transmission factor smaller than 1 but also because it can limit the light-gathering power : this limitation will be discussed in § 6.

5.2. *Exploration of the Spectrum*

From the formula $\sigma = K/2nt\cos i$, it can be seen that the displacement of the bandpass can be obtained by variation of one of the factors $\cos i$, K, n, or t. The most simple solution (as is used for the grating instrument) is to change the angle of incidence, either by the displacement of the diaphragm or by a rotation of the F.P. This solution has been used by Jaffe (1958) and by Humphreys and Paul (1958) ; its advantage lies in its simplicity, but it has the disadvantage of a large loss in luminosity imposed by the use of a slit instead of an annular diaphragm.

All other methods retain the revolution symmetry of the F.P. ; the latter is fixed while a circular diaphragm is centred on the normal to the F.P.

In the solution most often used the F.P. is placed in an airtight box inside which the pressure P is varied. The wavelength interval explored is equal to

$$\lambda_1 - \lambda_2 = \lambda(n_1 - n_2)/n_1 = \lambda C(P_1 - P_2)$$

C being a constant which depends on the gas used. For air it has the value $C = 3 \times 10^{-4}$ per atmosphere. For the study of hyperfine structure, the interval occupied by the structure is usually sufficiently narrow (about 1 Å) to be explored by a pressure variation of one atmosphere.

One thus simply evacuates the box and allows air to re-enter ; the only difficulty is to control the scanning rate. In the system of Kuhn and Ramsden (1958) the recording galvanometer spot is displaced in one direction by the signal, and in a direction at right angles by a bellows the expansion of which is proportional to the pressure. Other methods such as the supersonic leak of Rank and Shearer (1956), or the servo control system described by Dufour and Jacquinot (1948), and used by Blaise (1958) and Chantrel (1958) give directly a linear increase in air pressure. In the latter system the pressure is changed by an adjustable leak controlled by a mercury manometer which follows the displacement of a plunger lifted by a synchronous motor.

It is also possible to change the pressure by the displacement of a piston (Brody, Tomkins and Fred 1957, Baird 1958). For low resolution problems ($\mathscr{R} < 10^5$) the interval explored by a pressure change of one atmosphere becomes smaller than the free spectral range. As we shall see later, this prevents the exploration of a wide spectral domain. One has to use either a pressure variation greater than one atmosphere (Ring and Woolf 1958 : up to 100 atm) or another gas (Dupeyrat 1958 : freon). Jaffe (1958) has even suggested the uses of gases like CO_2 or Xe compressed up to the critical point where the index variations are very high : but it is probable that the density fluctuations would produce too great heterogeneity in the medium.

Other methods make use of a mechanical displacement of one of the F.P. plates. These methods have been chiefly developed for the domain of low resolutions, where the range which can be scanned by pressure variation is not sufficient. The difficulty is to keep the plates parallel to the necessary degree of accuracy while they are being moved : fortunately it is sufficient to displace the plate by $\lambda/2$ for scanning the free spectral range. Several solutions have been proposed. Dupeyrat (1958) has studied the use of piezoelectric spacers. Chabbal and Soulet (1958) use the elastic deformation of a metallic membrane in the centre of which is placed the plate. A mechanical force is applied to this membrane at two points suitably chosen. Kline (1956), Gobert (1958), Tolansky and Bradley (1959) and Greenler (1958) make use of the properties of the elastically deformable parallelogram. The results seem to be satisfactory with all these mechanical solutions provided the displacement does not exceed a few wavelengths.

However, it is still too early to decide which method will supersede the others for low resolution scanning.

None however appears to be capable of exploring continuously the wide spectral intervals usually encountered in practical spectroscopy. For this reason

Chabbal and Jacquinot (1955) have been led to introduce sawtooth scanning by means of which a spectral interval of unlimited extent may be explored, the only condition being that there exists a system suitable to explore one order. The principle is the following (figure 6). Let $\sigma_K, \sigma_{K+1}, \sigma_{K+2}, \ldots$ be the radiations transmitted

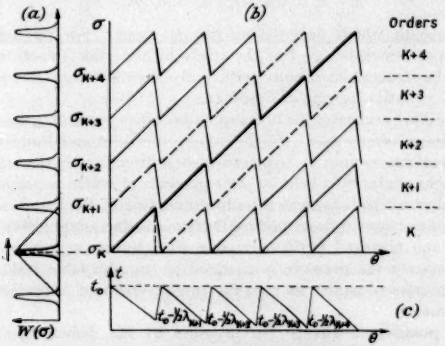


Figure 6. Sawtooth scanning (θ is the time).

by the F.P. in its initial state; with a pre-monochromator one isolates a band centred on σ_K. The optical thickness of the F.P. and the adjustment of the monochromator are changed in such a way that their passbands shift in synchronism up to the frequency σ_{K+1}. Now, without scanning the monochromator, the F.P. is returned abruptly to its initial thickness; it now transmits σ_{K+1} in the $(K+1)$th order. Once again one changes the F.P. and monochromator in synchronism until the frequency σ_{K+2} is reached; the F.P. is brought back to its initial position and the cycle starts again.

§ 6. The F.P. Spectrometer with a Single Bandpass

Until recently the F.P., in its spectrographic or spectroscopic form, has been employed almost only for hyperfine structure studies at high resolution. Its high luminosity at all resolutions has encouraged many workers to extend its use to the study of any spectroscopic problem, and in particular to that of absorption spectra. It is here that the shortcoming of the F.P., which is to transmit many radiations simultaneously, is vividly manifested. The exploration of a complex spectrum with such a multipeak instrumental function would yield results impossible to interpret; it is thus necessary to add to the F.P. a pre-monochromator capable of supplying the F.P. with a sufficiently narrow spectral interval. For the study of hyperfine structure it is sufficient to isolate the line under study, which rarely requires a resolvance above 5000. In the general case however the bandpass of the monochromator should not be greater than $\Delta\sigma$, so that all the transmission peaks of the F.P. are suppressed except one. This imposes on the monochromator a resolution $\mathscr{R}' = \sigma/\Delta\sigma$, only $1/N$th of that of the F.P. Furthermore it is desirable

that the light-gathering power of the monochromator be sufficient to illuminate fully the F.P. (i.e. both the focal diaphragm and the plates). If this condition is fulfilled, and if the transmission of the monochromator is high, the combination of F.P. plus monochromator will maintain its superiority of luminosity over the grating, and yet have the same range of applications.

Jaffe, Rank and Wiggins (1955) place a grating monochromator in front of the F.P.; the advantage of this solution is its simplicity, the disadvantage is the loss of light, which may be very serious. The optical system placed between the monochromator and the F.P. should image the grating on the isolating diaphragm, and the slit on the plates (or vice versa, but the result is the same). Suppose that the image of the grating covers exactly the isolating diaphragm; we have thus an image of the slit, of height at least equal to that of the plates, which has a width x times smaller than the width D of the plates. Then it may be shown that

$$x = MN(\sin i_1 + \sin i_2)/2D\sqrt{\mathscr{R}},$$

where M is the width of the grating.

For $D = 6$ cm, $M(\sin i_1 + \sin i_2) = 6$ cm and $N = 20$,

one obtains $x = 10/\sqrt{\mathscr{R}}$ equal to 1/10 for $\mathscr{R} = 10^4$ and 1/100 for $\mathscr{R} = 10^6$. In the last case x is still of the order 1/30 for the largest gratings commercially available. The gain of luminosity given by the F.P. may thus be lost; the only advantage is to increase the resolution above the $\mathscr{R}_0$ of any actual grating.

Hirschberg (1958) uses a grating monochromator in which the slits are portions of circles; their image can thus coincide with a part of the annular isolating diaphragm of the F.P. Here also, however, there is a loss in luminosity due to the fact that the slits may only occupy a fraction of the circumference and the F.P. remains only partially illuminated. This solution is chiefly interesting for low resolution work, where the slit width becomes greater than the aberrations; the slits may occupy, in the limiting case, one half of a circumference; the loss in solid angle is limited to π.

Chabbal (1957) has applied another solution which fulfils all the necessary conditions but which is more complex. One associates with the initial F.P. a second F.P. of smaller thickness the role of which is to eliminate the transmission peaks nearest to the wanted one. Since its resolvance is smaller than that of the first F.P. it does not limit the light-gathering power. On the other hand if the transmission is to remain high, and, at the same time, the unwanted peaks are to be weakened sufficiently, the free spectral interval cannot be made greater than about ten times that of the first F.P., an interval which is still in general much too small for most spectra. Thus one is obliged to add either a third F.P. or a grating monochromator of sufficient resolution to eliminate the unwanted peaks of the thinnest F.P.; in the latter case the slits can now be so wide that their image covers the whole surface of the F.P. A complete study of this problem has led to a set of formulae and curves, allowing the user to choose, for each particular problem, the best combination of Fabry–Perots and grating: the gain in luminosity over a single grating is widely variable, but remains in general very important; in the most unfavourable case this gain may be only slightly greater than one. Chabbal has constructed two spectrometers according to this principle, each formed by

the association of a grating monochromator with two F.P.; one is intended for the study of the visible spectrum at a resolution of about 10^4, the other (Chabbal and Noorman 1958) for near infra-red absorption work at a resolution of 5×10^5. Greenler (1958) has also constructed a spectrometer of this type to be used for the study of absorption spectra in the 10μ region.

The following points should be noted:

(1) When several monochromator elements are placed in series their bandpasses should shift simultaneously during the exploration of the spectrum. This can give rise to difficult problems but in the case of pressure scanning these are easily solved. The displacement of the bandpass of an interference device corresponding to a certain pressure change is the same, whatever the instrument; thus if all the elements are placed in the same chamber, they automatically keep in synchronism. Hirschberg, Kadesch and Mack (1958) have described an apparatus working on this principle. In the case where one wishes to use the pressure to adjust the bandpass of one F.P. with respect to another the adjustment remains valid if the *difference* in pressure between the two chambers remains constant. Scanning is thus possible if the pressures in the two chambers are varied in such a way that their difference remains constant (Chantrel 1958).

(2) Without going to the limit of a single bandpass, one may often wish to increase the finesse, that is to say the number of spectral elements contained in the free spectral range by placing in series two F.P. These may then be associated with a pre-monochromator of weak resolvance. This is the case of the double F.P. used by Brochard, Chantrel and Jacquinot (1958) in the study of the hyperfine structure of He_3, and by Lennuier in the study of the resonance line of mercury (see also § 9, double F.P. in photography).

§ 7. Technical Considerations

7.1. *Coatings*

The value of the transmission factor $\tau_A = [1 - A/(1-R)]^2$ and the value of the finesse $N_R = \pi\sqrt{R}/(1-R)$ depend on the properties of the coatings. Thus the improvement in coatings is an important factor in the general improvement in Fabry–Perot spectroscopy, although it is probably not the most important.

As shown by the expressions of τ_A and N_R, for a given value of the absorption coefficient A, an increase of R produces an increase of N_R but a decrease of τ_A. Thus the properties of a coating are best represented not by R and A, but by τ_A and N_R or by a point in the diagram $\tau_A = f(N_R)$. As already stated it is not necessary—and even not good—to have very high values of N_R; the aim should be to reach high values of τ_A for coatings having moderate values, say from 20 to 50, of N_R. In figure 7(*a*) are shown typical curves τ_A, N_R for conventional metallic films (Kuhn and Wilson 1950, Burridge, Kuhn and Pery 1953).

Great progress has been made with the introduction of multilayer dielectric coatings having very low absorption. A coating consisting of a small number of alternate high index and low index layers has a high reflecting power if the optical thickness of every layer is $\lambda/4$ (Cartwright and Turner 1939). This method is now well developed in the visible region (see for instance Dufour 1949, Giacomo 1955). Most often ZnS is used as the high index material and cryolite or MgF_2 as the low index material and five layers are sufficient to provide the reflecting power giving a

finesse N_R compatible with the actual N_D of the plates. The transmission τ_A is fairly good throughout the visible region as shown by the points in figure 7(*a*); at least below 6000 Å these coatings are far better than metallic films. The coatings are very stable, and do not show appreciable change after years. In the red and near infra-red regions, silver films have good performances but ZnS–MgF_2 coatings

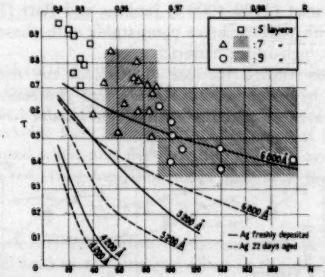


Figure 7(*a*). Silver coatings and dielectric coatings in the visible region (4000–6500 Å). Curves from Kuhn and Wilson (1950). Shaded zones and points: dielectric coatings ZnS–cryolite (Steudel 1958, Giacomo 1956).

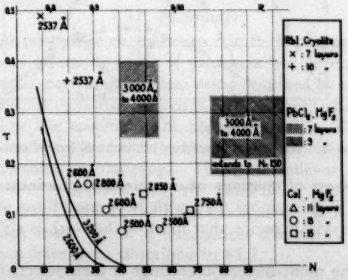


Figure 7(*b*). Aluminium coatings and dielectric coatings in the ultra-violet region.
Full curve: aluminium (Burridge, Kuhn and Pery 1953).
Shaded zones: dielectric coatings $PbCl_2$, MgF_2 (Steudel 1958).
Points: dielectric coatings CsI, MgF_2 (Steudel 1958), RbI cryolite (Lennuier 1958).

are very often used. For longer wavelengths semiconducting materials like Si, Ge, Te have very high indexes of refraction, of the order of 3 to 5, thus allowing good coatings to be made with only three layers, the low index materials being cryolite, MgF_2, NaCl, or KBr (Greenler 1958): these coatings may only be used above the limit of absorption of the semiconductor ($\lambda > 2\mu$ for Ge, $\lambda > 10\,\mu$ for Te).

The worst situation is in the ultra-violet region where, while the low index MgF_2 or cryolite layers are transparent to 2000 Å, it is very difficult to find high index materials which are still transparent. Steudel (1958) has obtained good

results above 3000 Å using $PbCl_2$: with seven layers at 3500 Å a finesse $N_R = 50$ can be easily obtained with $\tau_A \simeq 0{\cdot}3$. Between 2500 and 3000 Å he uses CsI as high index material: eleven layers are necessary to obtain $N_R = 25$ with τ_A only about 0·15. However the coatings made with CsI are very sensitive to humidity and must be kept in a dry atmosphere. Some results obtained by Steudel are shown in figure 7(*b*), and compared with those obtained with Al films.

In the near ultra-violet (3500–4000 Å) Jenkins and Barr (1956) have obtained very stable coatings with performances comparable with those of the ZnS–MgF_2 coatings with Sb_2O_3 as high-index material.

Lennuier (1958) has studied the use of RbI for the region around 2500 Å. This material is very hygroscopic, but an efficient protection of the coating is obtained with a $\lambda/2$ layer of paraffin. Although satisfactory results have been

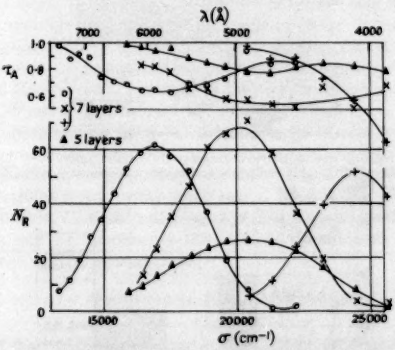


Figure 8. Chromatic properties of multi-dielectric coatings.

obtained with ten layers of coatings of cryolite–RbI, Lennuier does not claim to have reached better performances than with very good films of aluminium. It is certainly in this ultra-violet region that most of the work is still to be done on the coatings.

A drawback of the multi-dielectric coatings is their selectivity: figure 8 shows an example of the variation of N_R and τ_A for 5- and 7-layer coatings. Giacomo (1958) has pointed out that for wavelengths apart from the central region the ratio $(1-\tau_A)/N_R$† is approximately constant: that means that the coating is still ' good ' even if it is not employed in the region of its maximum finesse. However it would be preferable that both N_R and τ_A should change as little as possible in an extended domain. This problem has been considered by Penselin and Steudel (1955) and by Baumeister and Stone (1956); it is possible to extend the domain by varying the individual layer thicknesses; with 15 layers of alternate ZnS and cryolite of thicknesses increasing from 1000 Å to 1750 Å, Baumeister and Stone have obtained a coating the finesse of which stays greater than 40 from 4100 to 7200 Å.

† This relation proves only that the absorption $A = 1-T-R$ is a constant when the wavelength is changed: in effect $(1-\tau_A)/N_R \simeq 2.4/3$.

All these dielectric coatings are made by evaporation in a good vacuum. Of course the thicknesses must be controlled during the evaporation : that may be done either by continuously measuring the reflecting power with a galvanometer or with a more elaborate device (Giacomo and Jacquinot 1952) which measures the derivative of the reflecting power at the desired wavelength, or by measuring the reflecting power for two wavelengths symmetrical with respect to the desired wavelength.

7.2. *Materials*

The ideal material for F.P. plates is quartz or fused silica ; the useful spectral range with the best quartz now extends from 2000 Å to 3·5 μ. One can now easily obtain plates accurate to $\lambda_{green}/50$ over a diameter of 50 to 60 mm. The best plates to the author's knowledge have been made by Demarcq at the Institut d'Optique, Paris ; they are of fused silica, and flat to within $\lambda_{green}/70$ over a 140 mm diameter. For use in the infra-red, Ring has used plates of fluorite flat to within

$$\lambda_{green}/20 = \lambda(5\mu)/200,$$

and Greenler has used rock salt plates. It is reasonable to hope that the development of the F.P. will lead to a systematic effort in the production of high quality plates such as those of Demarcq.

It would thus be possible to take advantage of the progress made in producing coatings.

7.3. *Mountings*

The conventional procedure to make up a F.P. is to place between the two plates three spacers of quartz or invar of thickness not differing by more than a fraction of a wavelength, perfect parallelism being obtained by mechanical compression of the spacers. This mounting, which is certainly very stable, is often replaced by others which allow the spacing to be changed easily. This can be useful in hyperfine structure studies where the thickness has to be chosen critically in order to obtain less troublesome overlapping. Thus, at Bellevue (Dufour 1949), the plates are mounted each in a metal ring, the rings being separated by three silica spacers. The plates may be slid in the rings to adjust the separation to the desired value. Devices which make use of a mechanical displacement of one plate for scanning the spectrum have been described in § 5.2.

7.4. *General Arrangement*

The F.P. spectrometer is an ensemble which includes in series, the source, a monochromator, one or several F.P., an isolating diaphragm followed by the detector, and the scanning devices. In addition, between adjacent elements are placed optical systems designed to image the pupils of each element on the pupils of the following element. This optical system is often arranged to image the exit slit of the monochromator on the isolating diaphragm of the F.P. ; a better alternative is to image the F.P. on the slit of the monochromator : this allows the minimum surface of the F.P. to be employed (thus increasing the finesse) and allows the solid angle to be changed easily according to the resolvance required from the F.P. (Chabbal 1957).

The increased number of F.P. spectrometers constructed by various laboratories has led to the development of more compact and stable units, in which all elements are mounted on a single chassis. Such units are capable of operation in an ordinary laboratory environment.

§ 8. The Spherical Fabry–Perot

The fundamental law $L\mathscr{R} = \text{constant}$ makes the study of phenomena at a resolution of several millions difficult; not only does the luminosity of the F.P. itself fall at high resolution but the sources which one wishes to study at resolutions of this order are inherently weak. A new instrument invented by P. Connes (1956) and named the 'spherical Fabry–Perot', allows considerable progress in this domain.

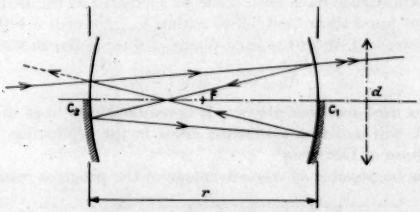


Figure 9. Spherical Fabry–Perot.

It consists (figure 9) of two spherical mirrors whose separation is equal to their radius of curvature r, so that paraxial foci coincide and the instrument is an afocal system. One half of the surface of the mirrors is semi-reflecting and the other half is fully reflecting. Any incident ray gives rise to an infinite number of outgoing rays which are *coincident* (and not only *parallel*, as with the plane F.P.). If one neglects aberrations their path difference is $\Delta = 4r$, a constant; but in order to do this one has to stop down both mirrors with circular diaphragms of diameter d; it can be shown that d increases with r, and that the light-gathering power is *proportional* to r (and thus to $\mathscr{R}_0$ instead of being *inversely proportional* to $\mathscr{R}_0$). For practically useful values of r (a few centimetres) d always remains quite small (a few millimetres), and the figuring of nearly perfect spherical plates of this size should be comparatively easy.

If one replaces a plane F.P. of thickness t and diameter D by the 'equivalent' spherical F.P. (thickness $r = t/2$, aberration-limited diameter d, and identical semi-reflecting coatings) one gets the same reflecting finesse N_R, the same free spectral range $\Delta\sigma = 1/4r = 1/2t$, the same theoretical resolving power $\mathscr{R}_0 = 4\pi N/\lambda$, similar (but not quite identical) instrumental functions, and a gain in light-gathering power $G = 2(t/d)^2 = (\mathscr{R}_0\lambda/2D)^2$. We see that G increases very rapidly when the thickness to be used is appreciably greater than the diameter of the plane F.P. with which comparison is made. Spectroscopic uses of the spherical F.P. are thus limited to the study of extremely sharp lines (given by atomic beams for instance). On the other hand the instrument may act as a monochromator yielding very sharp spectral bands with much larger flux than can be obtained by any other

means ; the coherence lengths of the corresponding wave trains could be very large (up to several hundred metres) and they could be used for fundamental experiments on the nature of light (Connes 1958).

§ 9. The Photographic Use of the F.P.

The small space which we devote to the photographic F.P. does not mean that its use has become less frequent. Actually many workers continue to use it in the conventional form for the study of hyperfine structure, but the technique has not made any significant advance.

We may recall however that the two dimensions of the photographic plate are best exploited if the dispersions of the F.P. and a grating are crossed. The slit of the monochromator is generally placed in the plane of the F.P. rings. In the case of a study of hyperfine structure the various spectral lines spread out along the x axis and are divided into horizontal arcs of the ring system which corresponds to the hyperfine structure of the line. The obvious advantage of this system is the simultaneous recording of the structure of many lines.

This method is equally applicable to absorption spectra, the only condition being that the bandpass of the monochromator be narrower than the free spectral range of the F.P.

The association of two F.P. has the result of an increased free spectral range as we have already seen, without lowering the resolvance. This technique has been used with success for instance by Jackson and Kahn (1938) and by Jackson (1956). The systematic errors which can be introduced in the measurement of absolute wavelength have been carefully considered by Jackson (1958).

Some workers have developed a technique whereby the image of the source coincides with the plane of the rings so that one can follow the changes of wavelength or spectral width at various points in the source. This has been applied in particular to astronomical problems.

Courtès (1951) has photographed the ring system given by the lines of nebulae in order to determine the relative Doppler displacements in different parts of the nebula. Jarrett and von Kluber (1955) use this method to measure the width of the green corona line in different regions of the solar corona. In another device described by Courtès (1958) the exit beam of the telescope is passed through a F.P. monochromator ; the resulting monochromatic radiation is focused on a photographic plate to give an image of the nebula. According to the resolution of the F.P. one can either isolate the line H_α, or neighbouring lines, or even parts of the H_α line which have different Doppler shifts ; one can thus draw a map of radial velocities over the nebula.

In the preceding experiments the pre-monochromator is an interference filter : if such a filter is inadequate a prism or grating monochromator must be used, with crossed dispersion. The region of the source which is thus observed is only that which is imaged on the entrance slit. Treanor (1957) has studied by this method the line shape of D_3 and H_α in the solar corona.

The photographic F.P. has also been used, at low resolution, for instance by Cabannes, and by Wark and Stone (1955) for the auroral lines.

It seems that up till now attention has not been drawn to the fact that the F.P. spectrograph is also more luminous than instruments of the same resolvance using

a prism or a grating. To obtain the best luminosity compatible with the resolvance it is necessary that the width of a ring be equal to the limit of resolution of the photographic plate. This condition may be satisfied by a suitable choice of the lens, the focal length of which should be $f = g\sqrt{(2m\mathscr{R}N)}$ where g is the grain size of the plate and m the rank of the ring used (starting from the centre). For instance for $m = 2$, $N = 20$, $g = 30\,\mu$, f should be 27 mm for $\mathscr{R} = 10^4$ and 270 mm for $\mathscr{R} = 10^6$. Under these optimum conditions, the illumination of the plate is

$$E = \tau S/2g^2 m\mathscr{R}N.$$

On the same basis the illumination given by a grating spectrograph (in Littrow mounting with blaze angle ϕ) would be $E = 4BS' \tan^2\phi/\mathscr{R}^2 g^2$. The gain G of the F.P. spectrograph over the grating spectrograph is

$$G = (S/S')(\mathscr{R}/8mN\tan^2\phi).$$

Assuming that $N = 20$, $\tan^2\phi = 1$, $S/S' = 0{\cdot}1$ and that the second ring is used ($m = 2$), the gain is of the order $\mathscr{R}/5000$: it increases with $\mathscr{R}$ and is greater than 1 even at moderate resolution. Of course the calculation implies that the focal length is properly adapted to the resolution, which can lead to very short focal lengths.

It is interesting to analyse why this gain increases with the resolvance. The gain in illumination is due to the reduction of the focal length, which is rendered possible by the increased dispersion (the basic rule being that the grain size and the width of the spectral image are equal). Now the dispersion of the F.P. approaches infinity at normal incidence and is very great on the rings of small angular diameter, that is to say for low rank and high resolution. If the resolution is decreased, the angular diameter of the rings increases and the dispersion may decrease to a value equal to or smaller than that of the grating.

In practice, in almost every domain of resolutions the F.P. is much more luminous. Its use in photographic instruments will no doubt become more frequent as in the case of recording spectrometers. Nevertheless it will probably always suffer from the fact that the dispersion is not linear.

PART III. METHODS MAKING USE OF THE AMPLITUDE OF INTERFEROMETRIC MODULATION

In this section we deal with methods in which the different spectral elements are resolved not by means of any spatial dispersion, but by means of an interferometric modulation. In these methods the *depth* of modulation is negligible for all elements except for those in a narrow spectral band. As far as we know only one of these methods has been fully developed, and it is so ingenious that it deserves a rather long section of this report. The instrument constructed by P. Connes (1958) is known as an interferometric spectrometer with selection by amplitude of modulation (abbreviation of the French name : SISAM).

§ 10. CONNES' METHOD

Before explaining the structure and the functioning of SISAM it is necessary to recall some properties of the Michelson interferometer used as an interferometric modulator.

10.1. *Michelson Interferometer*

A Michelson interferometer (figure 10), comprising two mirrors M_1, M_2, a dividing plate S, and a compensating plate C, is adjusted so that the image M_2' of M_2 is parallel to M_1. It is illuminated by a parallel beam of monochromatic light, of wave number σ_a coming from a point aperture D, and falling on the mirrors, at normal incidence. A condenser L concentrates the outgoing light flux on a detector R.

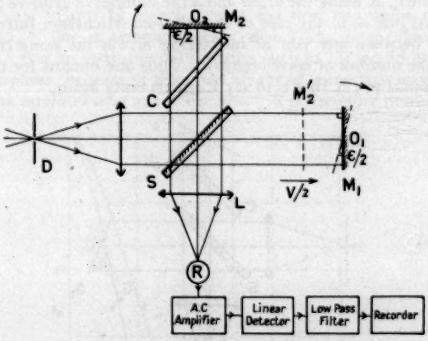


Figure 10. Michelson interferometer.

Under these conditions, the path difference Δ between two interfering rays is a constant, and the intensity of the transmitted flux can be expressed by

$$I = I_a \cos^2 \pi\sigma_a \Delta \propto I_a[1 + \cos 2\pi\sigma_a \Delta].$$

If Δ varies at speed V (for instance, by displacing one of the mirrors at speed $V/2$), that is to say if $\Delta = Vt$, the intensity becomes

$$I \propto I_a[1 + \cos 2\pi V\sigma_a t]. \qquad \ldots\ldots(1)$$

It is modulated at frequency $N = V\sigma_a$, and the depth of modulation P equals unity.

Now let us show the effect on P of a slight misadjustment of the interferometer. Suppose that the two mirrors rotate around two axes O_1, O_2 in their own planes, each by a small angle $\epsilon/2$. The two reflected beams are rotated through an angle ϵ and interfere at an angle of intersection 2ϵ : straight fringes appear on the mirrors ; these are parallel to the rotation axes and the fringe separation is $\lambda_a/2\epsilon = 1/2\epsilon\sigma_a$.

If the mirrors are rectangular, of width L, the new expression for the transmitted light is easily shown to be

$$I \propto I_a[1 + P\cos 2\pi V\sigma_a t] \quad \text{with} \quad P = \frac{\sin 2\pi\sigma_a L\epsilon}{2\pi\sigma_a L\epsilon} \qquad \ldots\ldots(2)$$

and the variation of P as a function of ϵ is shown by figure 9. One sees that P drops rapidly as ϵ increases ; its first zero corresponds to $\epsilon = 1/2L\sigma_a$, that is to say to a fringe separation equal to the width of the mirror. This curve can be directly recorded, with an a.c. amplifier, a linear detector and a low-pass filter.

10.2. *SISAM : Instrumental Function and Theoretical Resolving Power*

Now let us suppose that the mirrors are replaced by two identical dispersing elements (reflection gratings, prisms with a reflecting face or more complex combinations of prisms and gratings). Figure 11 shows the interferometer with two gratings G_1 and G_2 ; the image G_2' of G_2 intersects G_1 at its centre ; the angles of incidence (of the parallel beams considered hitherto) are equal to i_a and are chosen so that both beams are sent back exactly along their own paths (for σ_a only). This means that $\sigma_a = Kn/2 \sin i_a$, K being the order and n the number of grooves per centimetre. It is obvious that just as in the case of the adjusted Michelson interferometer, the path difference between any pair of interfering rays is the same (more exactly it differs by a whole number of wavelengths). Thus one obtains for the transmitted flux a value identical to (1), that is to say P equals unity again.

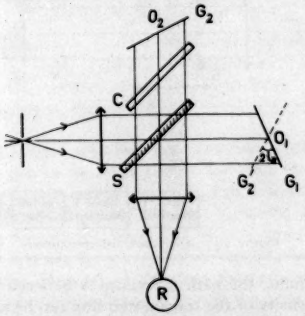


Figure 11. Principle of the SISAM.

But this is true for σ_a only ; for a neighbouring radiation of wave number σ, if both gratings keep the same orientation (that is to say if i_a remains the same) the two outgoing beams are rotated in opposite directions, each by an angle $\epsilon = 2 \tan i_a . (\sigma - \sigma_a)/\sigma$ and the depth of modulation drops ; it can be derived from expression (2) by replacing ϵ by its value as a function of the difference $\sigma - \sigma_a$ and by putting $\Delta_M = 2L \tan i_a$ (maximum value of the path difference) :

$$P = \frac{\sin 2\pi(\sigma - \sigma_a)\Delta_M}{2\pi(\sigma - \sigma_a)\Delta_M}.$$

Therefore the depth of modulation, which was equal to unity for $\sigma = \sigma_a$, drops very rapidly when σ deviates from σ_a ; only a narrow band of frequencies in the neighbourhood of σ_a is modulated and the instrument acts as a *selective modulator* for this band.

Let us now show how the interferometer gives a spectrum through direct recording just as a conventional spectrometer (with slits) would. Let us suppose the depth of modulation P to be recorded while the two gratings are rotated simultaneously with the *same* angular speed and in the *same* direction, R_1 and R_2' rotating then in *opposite* directions, so that the angles of incidence remain equal.

The wave number σ_a of the radiation to which the instrument is adjusted is now variable and scans the spectrum. A modulation peak is recorded each time that σ_a coincides with one of the lines. The instrumental function is, by definition, the recorded function if the spectrum is reduced to a single perfectly monochromatic line of wave number σ_0

$$W_0(\sigma_a - \sigma_0) = P = \frac{\sin 2\pi(\sigma_a - \sigma_0)\Delta_M}{2\pi(\sigma_a - \sigma_0)\Delta_M}. \qquad \ldots\ldots(3)$$

It can be seen (figure 12, curve a) that the first two zeros of W occur for $\sigma_a - \sigma_0 \simeq \pm 1/2\Delta_M$; one is tempted to say that $1/2\Delta_M$ is the limit of resolution of the system. But an instrumental function such as (3) would not be convenient for general purposes, because of the great height of secondary maxima, that would easily hide weak satellites near an intense line. It is therefore necessary to attenuate

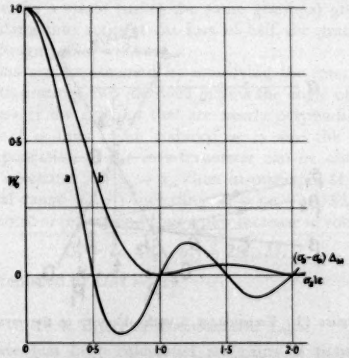


Figure 12. Instrumental functions of the SISAM without apodization (curve a) and with apodization (curve b).

them strongly. This operation, known as apodization (Jacquinot 1950), is sometimes made even with a slit spectrometer; in that case, it necessitates a variable density screen. The operation is here more necessary but relatively simpler, since the same result can be obtained with a diaphragm of a given shape. In the special case when this diaphragm is a diamond inscribed within the rectangle, one obtains

$$W_0(\sigma_a - \sigma_0) = \left(\frac{\sin \pi(\sigma_a - \sigma_0)\Delta_M}{\pi(\sigma_a - \sigma_0)\Delta_M}\right)^2$$

(figure 12, curve b), that is to say, an instrumental function *identical* with the one that is obtained by the use of a spectrometer with slits of negligible width, equipped with an *identical* grating (without diaphragm) in Littrow mounting. The theoretical resolution limit and the theoretical resolving power then take their well-known values $\delta\sigma_0 = 1/\Delta_M$ and $\mathscr{R}_0 = \Delta_M/\lambda_a$. Moreover, the use of the diamond is by no means necessary and it should often be possible to obtain a greater apodization

(or a smaller loss of grating area) with a diaphragm of a different shape. It can be shown, quite generally, that the instrumental line shape is the Fourier transform of the contour function of the diaphragm. But the diamond is very convenient for comparison with the conventional spectrometer, and it will therefore be assumed to be the diaphragm used.

10.3. *Permissible Solid Angle*

We have so far assumed the interferometer to be illuminated by a parallel beam of light, transmitting no energy; under these conditions its use would be of no practical value.

Let us now suppose the opening of the entrance diaphragm D to be finite; the interferometer receives not only rays such as α (figure 13) which are sent back upon their own paths, but rays such as β intersecting rays α at a small angle θ.

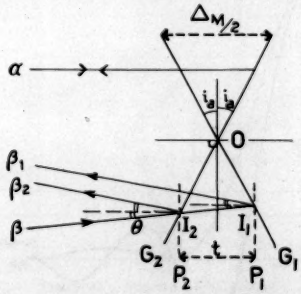


Figure 13. Variation of Δ with obliquity of the rays.

They can be skew rays, outside the plane of the figure; in any case they always give rise to two outgoing rays β_1 and β_2 that do not coincide but can be considered as parallel (neglecting terms of the order of θ^2); it is exactly as though β has been reflected in two plane mirrors P_1 and P_2, perpendicular to rays α and equivalent to a plane parallel plate of thickness t. Accordingly the path difference between β_1 and β_2 is $\Delta = 2t\cos\theta$ and its variation with θ is $\delta\Delta = -t\theta^2$, the maximum value of which is $(\delta\Delta)_M = -\Delta_M\theta^2/2$ (for rays incident on the marginal region of the gratings). This variation is exactly what one would get with a F.P. interferometer of thickness $\Delta_M/2$. Under these conditions it follows that a calculation similar to the one that has been made above in the case of the F.P. interferometer (and which we will not repeat) would lead to the following results:

(1) The aperture of the diaphragm D must be circular.

(2) The *actual* instrumental function W is the convolution of the *theoretical* function W_0 by a rectangular function whose width, f, is proportional to Ω, the solid angle subtended by D.

(3) The optimum value of f is $f = \delta\sigma_0$; the widening of W relative to W_0 and the lowering of the actual resolving power $\mathscr{R}$ relative to $\mathscr{R}_0$ being then such

as to give $\mathscr{R} = 0{\cdot}8\mathscr{R}_0$. The corresponding value of Ω is again given by the fundamental relation

$$\Omega\mathscr{R}_0 = 2\pi. \qquad \ldots\ldots(4)$$

(4) It is useless to try widening Ω any further because the modulated flux increases very little while $\mathscr{R}$ drops rapidly. In consequence the SISAM is for all practical purposes an instrument of *fixed* resolving power ; in order to change the value of $\mathscr{R}$ it is necessary to vary the angle of incidence on both gratings, and this in turn calls for gratings of different blaze angle.

In short, while the calculation of $\mathscr{R}_0$ closely resembles the same calculation for the slit spectrometer, that of Ω is nearly identical with that of the F.P. ; this shows perfectly the hybrid nature of the SISAM which to a certain extent combines the essential qualities of these two instruments.

From a practical point of view it is interesting to notice that the replacement of a slit spectrometer by a SISAM (using the same gratings) gives a gain in light-gathering power (taking into account the loss of half the grating surface, due to the apodizing diaphragm) $G = \pi/\beta \tan i_a$.

Much larger gains can be obtained by modifying the interferometer. Let us place in front of both gratings two identical prisms the angle of which is so chosen as to give two images of the gratings that are nearly perpendicular to ray α ; the equivalent thickness t is thus much reduced as is also the variation $\delta\Delta = t\theta^2$. Nearly perfect compensation of the interferometer can be obtained for only one radiation σ_c on the spectrum ; if $\sigma_a = \sigma_c$, then in principle $\Omega = 2\pi$. In order to scan a given spectral range (σ_2, σ_1) including $M = (\sigma_2 - \sigma_1)/\delta\sigma_0$ spectral elements, σ_c is chosen in the centre of (σ_2, σ_1) and thus a new increase in solid angle is obtained,

$$G' = 2\sigma_c/(\sigma_2 - \sigma_1) = 2\mathscr{R}_0/M. \qquad \ldots\ldots(5)$$

Formula (4) is then replaced by $\Omega M = 4\pi$.

10.4. *Effect of Unmodulated Light*

Gains which have just been calculated are gains of light-gathering power ; but since the SISAM (unlike for example the F.P. interferometer) is a selective modulator and not a monochromator, radiations near the one that is being modulated at a given time can reach the detector and their effects must be considered. There are two different cases :

(*a*) Noise from thermal or photoconductive detectors (which are the only available ones throughout most of the infra-red) is strictly *detector noise*, that is to say it is for all practical purposes unmodified by any reasonable amount of unmodulated light, the action of which can thus be neglected. The available signal is then multiplied by G but at the same time it is, in principle, necessary to multiply by the same factor the area of the receiver, the noise of which is thus multiplied by $\sqrt{G}$; then the gain in signal to noise ratio is $G_{s/n} = \sqrt{G}$.

(*b*) On the other hand, noise from photo-emissive detectors (useful in the visible ultra-violet and near infra-red) is *photon* noise, proportional to the square root of total incident flux ; it is strongly increased by the unmodulated part of the light, and this robs the SISAM of much of its advantage ; however in the case of spectra with small average density (such as Raman spectra for instance) the gain can still be of interest.

10.5. *Practical Considerations on the Use of SISAM*

It is not convenient to vary the path difference by displacing one of the gratings ; a rotation of the compensating plate gives the same result with much greater ease. As this rotation cannot be continuous a symmetrical sawtooth motion is used ; this means that Δ varies alternately with the speed V or $-V$. The modulated signal consists in incoherent wave trains, each containing n waves. This is of no practical importance, provided n stays much greater than the Q factor of the band-pass filter which is obviously necessary in the a.c. amplifier (in order to reduce the noise). Moreover this filter plays another part. Since the modulation frequency is proportional to the light frequency, it eliminates modulated signals of unwanted grating orders, thus replacing an optical pre-monochromator (prism or filter).

All optical surfaces within the interferometer (gratings, dividing and compensating plates) must approach theoretical perfection ; the errors have to be small compared with the wavelength. This condition is fortunately not too difficult to fulfil in the infra-red. The same can be said of the necessary mechanical stability of the interferometer, and chiefly of the mechanical precision of the rotation axes. On the other hand imperfections of optical surfaces outside the interferometer matter very little and there is no difficulty in keeping aberrations negligibly small.

10.6. *First Results*

Two interferometers have been built so far ; the first is a small grating instrument with a limit of resolution of 0·2 cm^{-1} ; the second one is of the compensated prism and grating type. Both are intended for work in the near infra-red (up to 3 μ) and have been shown to give the calculated resolving power and the calculated increase in light-gathering power.

§ 11. Guilino and Maier-Leibnitz' Method

Guilino and Maier-Leibnitz (1958) describe another type of selective modulator which they have built for recording night sky emission spectra. A Fabry–Perot interferometer (figure 14) of thickness t is used in reflection at about 45° incidence.

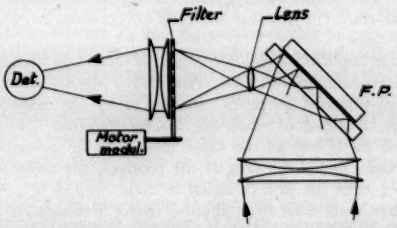


Figure 14. Guilino and Maier–Leibnitz' spectrometer.

In the focal plane the fringes are small portions of ellipses and can be very roughly considered as straight and equidistant ; they are imaged in the focal plane of a lens. First a grid is made by photographing the fringes of a monochromatic source

of wavelength λ_1, then the developed photograph is put back in its place, and the fringes of the polychromatic source to be studied are imaged on it. Modulation is achieved by giving the grid an oscillating motion the amplitude of which is one quarter of the average fringe separation. It is clear that only the light of wavelength λ_1 is fully modulated ; for nearby lines modulation decreases rapidly (because fringes cannot coincide with grid lines over the whole field of view). Resolving power is approximately equal to the number of grid lines and can only be a fraction of the theoretical resolving power of the interferometer. Scanning is achieved by varying the thickness of the interferometer, thus changing the fringe separation ; it can easily be shown that if thickness becomes t the instrument is adjusted (i.e. gives full modulation) for wavelength $\lambda = \lambda_1 t/t_1$. The factors limiting the solid angle are not clear and comparison with other methods is thus made difficult. No spectra recorded by this method seem to have been published as yet.

PART IV. METHODS MAKING USE OF THE FREQUENCY OF INTERFEROMETRIC MODULATION. SPECTROSCOPY BY FOURIER TRANSFORMATION

§ 12. General Principles of the Method

In the past few years a new and very promising form of spectroscopy has made rapid progress. This method, which is very different from other methods, is based on the following fundamental theorem :

Theorem. The flux which emerges from a two-beam interferometer illuminated by a polychromatic source is the sum of two terms of which one, ϕ_0, is independent of the path difference, while the other, ϕ', is the Fourier transform of the spectrum of the source.

Let us first assume that the source is monochromatic with a wave number σ and a luminance $B(\sigma)$. The emerging flux, for parallel rays, is

$$\phi \propto B(\sigma)\cos^2\pi\sigma\Delta \propto B(\sigma) + B(\sigma)\cos 2\pi\sigma\Delta$$

where Δ is the path difference between the two beams. We are interested only in the variable part of the flux.

If the source consists of polychromatic light, with a luminance $B(\sigma)\,d\sigma$ in the region $(\sigma, \sigma + d\sigma)$, every spectral region $d\sigma$ makes a contribution to the variable flux

$$d\phi' \propto B(\sigma)\,d\sigma\cos 2\pi\sigma\Delta$$

and the total flux is

$$\phi(\Delta) = \phi_0 + \phi'(\Delta) \propto \int_0^\infty B(\sigma)\,d\sigma + \int_0^\infty B(\sigma)\cos 2\pi\sigma\Delta\,d\sigma.$$

The second integral, ϕ', is the Fourier cosine transform of the spectral distribution of the light source $B(\sigma)$.

Since the Fourier transformation is reciprocal, the spectrum $B(\sigma)$ can be obtained from the variation of the flux, $\phi'(\Delta)$, by means of an analogous relation

$$B(\sigma) \propto \int_0^\infty \phi'(\Delta)\cos 2\pi\sigma\Delta\,d\Delta. \qquad \ldots\ldots(6)$$

If the path difference of a two-beam interferometer is made to change progressively, the emerging signal $\phi'(\Delta)$ can be registered by means of a detector

(a photoelectric cell, a photoconductive cell, etc.); it is called ' the interferogram function ', and to obtain the required spectrum $B(\sigma)$ it is only necessary to find the Fourier cosine transform of the interferogram function.

This procedure is related to Michelson's earlier method of ' visibility of fringes '. Nevertheless it differs from this essentially; the relationships between the two methods are analysed below (§ 18).

The first, very crude, application of this method was made in 1911 by Rubens and Wood. The remarkable properties and the advantages of ' spectroscopy by Fourier transforms ' were first brought to light by Fellgett (1952, 1958) and by Jacquinot (1954 a, 1958).

§ 13. Resolving Power

As in all spectroscopic methods, the resolving power is determined by the maximum path difference between two rays between which there is interference. To find the resolving power we must consider the instrumental function, obtained when the instrument is illuminated with monochromatic radiation, $B(\sigma) = \delta(\sigma - \sigma_0)$. Since the extent to which the movable element of the interferometer can be displaced is limited, the path difference cannot exceed a maximum value Δ_M. The interferogram function $\phi'(\Delta)$ is therefore $\phi'(\Delta) = \cos 2\pi\sigma_0\Delta$ for values of Δ between 0 and Δ_M, and $\phi'(\Delta) = 0$ for values of Δ exceeding this. The same result is obtained if Δ varies between $-\Delta_M$ and $+\Delta_M$ since $\phi'(-\Delta) = \phi'(\Delta)$. The instrumental function is given by integral (6) with limits of integration 0 to Δ_M, instead of 0 to ∞. From this it follows that

$$W(\sigma) \propto \int_0^{\Delta_M} \cos 2\pi\sigma_0\Delta \cos 2\pi\sigma\Delta \, d\Delta \ast B(\sigma)$$

and from this, the instrumental function

$$W(\sigma) \propto \frac{\sin 2\pi(\sigma-\sigma_0)\Delta_M}{2\pi(\sigma-\sigma_0)\Delta_M} + \frac{\sin 2\pi(\sigma+\sigma_0)\Delta_M}{2\pi(\sigma+\sigma_0)\Delta_M}. \qquad \ldots\ldots(7)$$

The second term, identical with the first if σ_0 is changed to $-\sigma_0$, is an instrumental function centred on a negative frequency and plays no part if $\Delta_M \gg \lambda$, which is always the case. We can therefore take as the instrumental function

$$W(\sigma-\sigma_0) = \frac{\sin 2\pi(\sigma-\sigma_0)\Delta_M}{2\pi(\sigma-\sigma_0)\Delta_M}.$$

This is the same function as (3), the first two zeros of which correspond to

$$\sigma - \sigma_0 = \pm 1/2\Delta_M.$$

Its ' width ' is $\delta\sigma = 1/2\Delta_M$ and the resolving power is $\mathscr{R}_0 = 2\Delta_M/\lambda$. It is represented by curve a, figure 12.

It is apparent that, as in the case of a SISAM without arrangements for apodization, the instrumental function is identical with an amplitude diffraction function, with negative components and high values of the secondary maxima. For this reason it is advantageous to introduce an apodization function $T(\Delta)$ so that the function becomes

$$W(\sigma-\sigma_0) = \int_0^{\Delta_M} T(\Delta) \cos 2\pi\sigma_0\Delta \cos 2\pi\sigma\Delta \, d\Delta.$$

For example, it is found that if $T(\Delta) = 1 - \Delta/\Delta_M$ for $0 < \Delta < \Delta_M$ then

$$W(\sigma - \sigma_0) = \left(\frac{\sin \pi(\sigma - \sigma_0)\Delta_M}{\pi(\sigma - \sigma_0)\Delta_M}\right)^2 \quad \text{(figure 12, curve b)}.$$

We now have an intensity diffraction function, whose first two zeros correspond to $\sigma - \sigma_0 = \pm 1/\Delta_M$. The width w of the instrumental function is greater and the resolving power is $\mathscr{R} = \Delta_M/\lambda$. The form of the instrumental function is thus identical with that of other types of apparatus, and is much more favourable than the preceding one. Here the same general relations between $\mathscr{R}$ and Δ_M are found as for other methods.

It is again evident that other forms of $T(\Delta)$ can be found, giving for the same value of Δ_M a rather higher resolving power, while retaining a favourable form of the instrumental function (Connes and Gush 1959, Gebbie 1959).

The physical procedure by means of which the apodization function is applied is quite different from that in the case of conventional spectrographs or the SISAM. In principle, it would be possible to weaken the light falling on the receiver in the ratio $T(\Delta)$ as a function of the path difference Δ, but this would give a very poor solution, producing a deterioration of the signal to noise ratio produced by the detector. It is much better to produce, following the function $T(\Delta)$, a weakening of the signal given by the detector, since the noise is reduced simultaneously, and the same signal to noise ratio is retained. In fact this reduction can be effected according to any required law on the interferogram function $\phi'(\Delta)$ itself, before calculating the Fourier transform. This point will be considered in § 16.3.

§ 14. Admissible Solid Angle

In the theory developed above, it has been assumed that the path difference is identical for all the rays entering the detector. But if a finite amount of energy is to reach the detector, the solid angle of the rays must be finite, and it is necessary to accept rays which have differing path differences : it is important to find out how much the resolving power is reduced as the solid angle is increased. This calculation can be made only if the type of interferometer is defined ; a general calculation is not possible. In this paragraph the calculation will be made only for interferometers of the Michelson type. The calculation is therefore very similar to that made for the Fabry–Perot interferometer and the SISAM, and the results obtained in § 5.1 and § 10.3 can be used. In effect, the path difference Δ depends on the angle of incidence, according to the same relationship, $\Delta = \Delta_0 \cos i$, where Δ_0 is the path difference for normal incidence ; on the other hand, the resolving power depends on the maximum path difference Δ_M according to the relationship $\mathscr{R}_0 = \Delta_M/\lambda$. The relationship between the admissible solid angle Ω and the resolving power $\mathscr{R}_0$ is therefore the same $\Omega\mathscr{R}_0 = 2\pi$. All the advantages associated with this high value of the product $\Omega\mathscr{R}$, which have already been described in the case of the Fabry–Perot interferometer and the SISAM, are therefore valid also for systems in which the Michelson interferometer is used. For other systems, the product can be much smaller.

It is possible further to increase the product $\Omega\mathscr{R}_0$ beyond the limit 2π by use of the systems described by Connes (1956) and Mertz (1958). The principle,

which is common to these systems and which can be compared with that of the spherical Fabry–Perot interferometer or with the compensating prisms systems used with the SISAM, depends on taking from a single incident ray, emergent rays which are coincident and not only parallel. For this purpose, Connes uses an afocal system situated on each of the arms of a Michelson interferometer ; Mertz makes use of one plate with parallel surfaces and of constant thickness, and a second plate whose thickness varies simultaneously with the path difference.

§ 15. Fundamental Advantage of the Method

In photographic methods, the spectrum is split up into its different elements spatially, and all the spectral elements investigated affect simultaneously the different parts of the photographic plate. The time T required to photograph all the spectral elements, which may number many thousands, is the same as that required for a single element, and all the available information is exploited to the greatest advantage. On the other hand, in the methods depending on registering signals, a single element only is accepted at any instant by the detector ; the time T' required for registering this element, determined in general by the signal to noise ratio, increases as the light source becomes weaker, the light-gathering power of the instrument becomes smaller and the detector becomes less efficient ; to register M elements, the time required is MT', and may become considerable for extensive studies, although the efficiency of the photoelectric cell, which is greater than that of the photographic plate, gives a value of T' smaller than that of T, for any one element. This inferiority can be partly palliated by using several channels, but their number must obviously remain relatively very small while the photographic plate behaves as a system with a number of channels equal to the number M of the elements to be investigated. Unfortunately there exists no photographic plate (or any analogous arrangement) which can be used for wavelengths longer than 1 μ. In the method which we are now studying, no information is rejected, all spectral elements affecting the detector simultaneously and, as with the photographic method, all the available information being exploited to the maximum. The time T' required for observing a single spectral element is determined, as with other methods, by the quantity of energy received from each spectral element by the receiver in the time of the run 0–Δ_M ; T' is therefore the same as for any other spectrometer with the same light-gathering power for equal resolving power, but the time required to study M elements is also T' instead of being MT', as it is with other methods, and it is evident that the gain is substantial, above all for investigations over extended regions. This system also behaves as a system whose number of channels is equal to the number of spectral elements to be investigated. As suggested by Fellgett (1958), the system can be regarded as a multiplex spectrometer, in which the different spectral elements are distinguished from each other by sinusoidal modulation of different frequencies, which are orthogonal, and can be separated by filters or by Fourier analysis.

However, the fundamental advantage of the Fourier method disappears if the effect of noise spoils the independence of the effects of the different spectral elements on the detector. The arguments concerning the effects of noise are the same as in the case of the SISAM (§ 10.4) and need not be repeated. When the detector noise is negligible compared with the photon noise, the noise due to M spectral

elements adds up quadratically, the signal to noise ratio is divided by $\sqrt{M}$,† and the time required for investigating each spectral element becomes MT' instead of T' : the advantage of the method is therefore lost completely.

The Fourier method is at its most efficient in the infra-red ; it is most useful in this region on account of the weakness of the light sources and the poor efficiency of detectors.

§ 16. Transformation of the Interferogram into the Spectrum

A Fourier analysis of the function $\phi'(\Delta)$ is an operation which is necessary for the reconstruction of the spectrum $B(\sigma)$ (or, more precisely, the spectrum deformed by the instrument $B'(\sigma) = B(\sigma) \star W(\sigma)$) from the interferogram. There are several methods of performing this operation : in the first two, the data concerning the spectrum are provided immediately, while in the others a recording is made and this is transformed later.

16.1. *Direct Use of Electric Filters*

If the path difference is a linear function of time (i.e. if the movable element of the interferometer is displaced at an exactly constant rate), then $\Delta = Vt$ and the interferogram given by a monochromatic radiation is $\phi'(t) = a \cos 2\pi\sigma_0 Vt$. The electric signal provided by the detector has a frequency $F = \sigma_0 V = \nu_0 V/c$, proportional to ν_0, the frequency of the radiation, and it can be isolated with an electric filter. The analysis of the complex signal given by a polychromatic source could therefore be effected by feeding, in parallel, m filters, tuned to different frequencies. The apparatus would thus have facilities for analysis as a polychromator with m channels. It is apparent that the selectivity of each channel would have to be at least as great as the resolving power, $\mathscr{R}_0 = \Delta_M/\lambda$, in order to avoid reducing the resolving power of the system. The number m of the channels could never be very large, and the possibilities of the method described in § 15 would consequently not be well exploited. Nevertheless, an arrangement of this type would have the advantage that it would give the results immediately, and it could be used for measuring the relative intensities of a number of pre-selected spectral elements in the spectrum of a weak light source.

The procedure for making measurements would be to make a single run with the mobile element of the interferometer and to read, at the end of this operation, the indications given by the different channels. If the selectivity Q of each channel is precisely equal to the resolving power, the time constant τ is equal to $Q/\pi F$ or $\Delta_M/\pi\lambda F$; it is determined by the selectivity and is, indeed, of the same order of magnitude as the time $T = \Delta_M/\lambda F$ required for a single run of the interferometer. Up to the time of writing, this method does not appear to have been put into practice.

16.2. *Instantaneous Analogue Computer of Strong and Vanasse*

The instantaneous analogue computer (Strong and Vanasse 1958) is related to the system described in the preceding paragraph : the selective filters are replaced by ten channels, each providing synchronous rectification and mechanical

† This is true only if all the spectral elements are of comparable intensity.

integration. The synchronous rectification is provided in each channel by a rectangular signal, produced by a mechanical interrupter, which is controlled by the mechanism associated with the mobile element of the interferometer. By means of suitable gears, ten reference frequencies are given neighbouring values which enable ten neighbouring spectral elements to be measured with a resolving power of 100. The integration is effected by a sphere and disc system, and is read off from a numerical counter. Ten consecutive spectral elements are thus obtained. To investigate another region of the spectrum the rectification frequencies F are kept unchanged, but the speed V is altered, which changes the values of σ since $\sigma = F/V$.

This system has the same advantages and shortcomings as that described in the preceding paragraph. However, the integration and filtering systems are better, since the various channels are given automatically appropriate frequencies and selectivities. Nevertheless, it seems to have been rejected by its authors, since Vanasse, Strong and Loewenstein (1959) have subsequently published new results obtained by means of a numerical transformation.

16.3. *Numerical Transformation*

This method is non-instantaneous, and the interferogram function $\phi'(\Delta)$ is registered either in the form of a curve produced by a pen recorder, or in numerical form for a certain number N of equally spaced values of Δ, $\Delta_n = n\Delta_e$. In the former case, N values of $\phi'(\Delta)$ are measured on the curve at the values Δ_n of Δ. Numerical calculation of the Fourier transform is made subsequently by replacing the integral by a sum of N discrete values of $\phi'(\Delta)$

$$B''(\sigma) = \sum_{0}^{N} \phi'(\Delta_n) \cos 2\pi\sigma\Delta_n.$$

It is, in general, not necessary to calculate $B''(\sigma)$ for values of σ closer than the limit of resolution $\delta\sigma = 1/\Delta_M$. If the spectral range to be studied extends from σ_A to σ_B, the number of summations to be made is equal to the number of spectral elements contained in this range, that is $M = (\sigma_A - \sigma_B)/\delta\sigma$.

The spectrum $B''(\sigma)$ found by this means is not identical with $B'(\sigma)$, but it can easily be shown that it is derived from $B'(\sigma)$ by a periodic reproduction at wave number intervals $\Delta\sigma = 1/\Delta_e$ and thus

$$B''(\sigma) = \sum_{-\infty}^{+\infty} B'(\sigma - n\Delta\sigma),$$

where n is an integer. It is apparent that the superposition of all these spectra becomes more troublesome as the repetition interval becomes smaller. The interval $\Delta\sigma$ plays precisely the same part as the free spectral range of the Fabry–Perot interferometer ; as with the latter, $\Delta\sigma = 1/\Delta_e$, i.e. it is inversely proportional to the path difference Δ_e. Similarly $\Delta_e = \Delta_M/N = 1/N\,\delta\sigma$ and thus $\Delta\sigma = N\,\delta\sigma$. The Fourier transformation method, in which N discrete, equidistant values of the path difference of the interferometer are used, is therefore equivalent to a Fabry–Perot interferometer with a finesse N, or a grating with N lines.

In reality a more complete analysis should take into account the negative values of σ due to the second term, in equation (7), of the instrument function, which

was neglected in § 13. It follows, therefore, that

$$B''(\sigma) = \sum_{-\infty}^{+\infty} B'(\pm\sigma - n\Delta\sigma),$$

and the *maximum*† value of $\Delta\sigma$ is equal to $1/2\Delta_x$, and not $1/\Delta_x$. The *minimum* number of samplings of the interferogram which must be made in order to obtain a free spectral range $\Delta\sigma$ is therefore given by $N = 2\Delta\sigma/\delta\sigma$ instead of $N = \Delta\sigma/\delta\sigma$. Thus, the maximum number of spectral elements which can be determined without overlapping is $M = N/2$. If the spectral range covered by the light source is narrower than $\Delta\sigma$ there is no overlapping; if this is not the case, it is necessary to isolate a range smaller than $\Delta\sigma$ in this spectrum. This can be effected by means of a light filter or an optical monochromator; but the same result can be obtained more efficiently by means of an electric bandpass filter. In fact, if the path difference is varied according to the law $\Delta = Vt$, the isolation in the signal of a frequency band ΔF is equivalent to isolating a spectral range $\Delta\sigma$, equal to $\Delta F/V$. There are two advantages in using an electric filter instead of an optical filter. First, an electric filter does not weaken the incident light; secondly, in addition to removing unwanted frequency bands, it also reduces the noise of the detector by eliminating the noise due to all of the spectrum outside the range passed by the filter.

A particularly elegant method of obtaining the filtering described above is that proposed by Mertz (1959), under the name of 'heterodyne spectrometry'. It depends on imposing a change of frequency on the signal given by the detector by means of a heterodyne signal produced by a portion of the interferometer illuminated by a monochromatic reference line. If the reference line coincides with one of the limits σ_A of the spectral range $\sigma_A - \sigma_B$ which is to be investigated, the frequency spectrum covered by the complex signal resulting from the frequency change extends from frequency 0 to a frequency F_B equal to $(\sigma_B - \sigma_A)\,V/C$: for filtering, therefore, a low-pass filter (which can be the recorder itself) can be used instead of a bandpass filter. To avoid any ambiguity due to image frequencies it is necessary to use two heterodyne signals in quadrature, and to make two recordings, of which one subsequently adds the Fourier transforms.

The numerical calculation is effected by means of digital computers such as the EDSAC (Fellgett 1958, Gebbie 1958) or the IBM 650 (Vanasse, Strong and Loewenstein 1959) or the IBM 704 (Connes and Gush 1959). With the possibilities of the machines and the methods of programming now available the machine time required for making each summation

$$\sum_{0}^{N} \phi'(\Delta_n) \cos 2\pi\sigma\Delta_n$$

is of the order of N milliseconds. As we have seen, the minimum number of sampling points necessary for the determination of M spectral elements without overlapping is $N = 2M$; thus the machine time required for the computation of these M elements is $MN = 2M^2$ milliseconds. This time can become quite considerable in the case of spectra covering a large range, measured with high resolving power. (For example, of the order of two hours for a spectrum with

† *Note added in proof.* A recent article (Jacquinot 1960) gives a more detailed treatment of this particular point.

2000 spectral elements.) Since the Fourier method is of the greatest value in the case of spectra with a large number of spectral elements, it is evident that the machine time is an important consideration.

An apodization factor $T(\Delta)$ can be introduced without difficulty (Gebbie 1958): it is sufficient to programme the machine to give for every value of $\phi'(\Delta)$ multiplication not only by $\cos 2\pi\sigma\Delta_n$ but also by $T(\Delta_n)$, the function $T(\Delta)$ being given any desired shape. It is apparent that this operation is effected without any loss of information (which is not the case with optical apodization).

It is of the greatest importance that there be no error in the values of Δ_n due either to irregularity in the rate of displacement of the movable element of the interferometer, or to distortion of the abscissa scale of the interferogram. Errors of these types would produce effects similar to those produced by errors in the ruling of a grating. Moreover it may be shown (J. Connes 1958) that the effects are much more serious, since they are proportional to the magnitude of the error, instead of to its square. Since the interferometers now in use are mechanically much less perfect than grating ruling engines there is a serious risk of getting erroneous results (ghosts and intense spurious lines) if these are based entirely on the measurement of equidistant ordinates in the interferogram. Consequently it is necessary to record on the interferogram a calibration which is independent of all these mechanical defects. This can be achieved by recording, on the same paper as the interferogram, fringes of a monochromatic reference line. There is no reason why the reference line should not be in the visible region even if the spectrum which is being investigated is in the infra-red. If the values of $\phi'(\Delta_n)$ are given in a numerical form by means of a digitizer, the functioning of the digitizer can be controlled by the fringes due to the reference line (Gebbie 1959).

16.4. *Application of Specialized Computers*

In order to render the Fourier method completely practical, the spectrum should be able to be reconstituted immediately after the production of the interferogram; this cannot easily be realized by means of numerical transformation, since this necessitates the use of large, modern electronic computers. There is certainly the possibility of constructing a specialized computer much less costly and complicated than digital computers which are not specialized. P. Connes is now engaged in investigating, in the Laboratoires de Bellevue, a very promising means of achieving this end. In the apparatus which he is developing the interferogram will be recorded on a magnetic tape in the form of a closed loop simultaneously with the reference signals provided by a monochromatic line. During the course of the analysis, the loop turns continuously, one spectral element being measured by one rotation of the loop. This analysis is effected by multiplying the recorded signal by a sinusoidal signal derived from the reference signal by frequency multiplication in the ratio p/q; p and q can have very high values and consequently there is no limit to the closeness of the points in the spectrum to be investigated. The signal obtained after multiplication is integrated in such a manner that the Fourier integral is obtained for every point which is analysed. In the spectrum which is obtained finally, the abscissae are thus determined with high precision by a digital process while the ordinates are obtained by an analogue process. With this method, there will be no overlapping, just as with a prism.

Since the machine is only in course of construction it is not yet possible to compare this method with the method of numerical transformation, which has already been tested.

§ 17. Experimental Results

In the first experiment, made by Rubens and Wood (1911), the interferometer consisted of two parallel plates of quartz, without metallic coatings, which gave two-beam Haidinger fringes. The analysis was very crude, and the only conclusion which could be made was that the light source contained radiations of wavelengths 150 and 200 μ.

Fellgett (1958), working in the region of 2 μ, has used a Michelson interferometer, in which the plane mirrors were replaced by cube corner reflectors; a double-sided reflecting chopper associated with two lead sulphide detectors mounted in opposition enables the two beams of the interferometer to be utilized. One of the corner cubes is displaced by means of a lever reduction gear at a speed of 60 wavelengths per minute. The apparatus was set up in the Coudé focus of the 35 inch telescope at the Cambridge observatory. Spectra of stars down to the fourth and fifth magnitudes have been obtained with a resolving power of 60. In the preliminary experiments the Fourier transformations were made numerically on Lipson–Beevers tapes.

A Michelson interferometer was used also by Gebbie in some recent experiments (1959) in the far infra-red (~ 200 μ) carried out at the National Physical Laboratory. The interferometer is a conventional Michelson interferometer, with a beam splitter consisting of a stretched membrane of polythene, 20 μ thick, coated with a layer of germanium. The travel available is 2·5 cm, giving a theoretical limit of resolution of 0·4 cm^{-1}. The recording time was about 30 minutes. The interferogram was sampled at 528 equally spaced intervals and the data automatically printed on punched tape. Using the National Physical Laboratory DEUCE computer, the time required to compute the spectrum over the range 125 to 225 μ was 5 minutes.

A completely conventional Michelson interferometer was also used by J. Connes (1958) in her first experiments, and more recently, by J. Connes and Gush, in the region 1 to 2 μ. In the latter experiment, the spectrum of the night sky was observed with a resolving power of 1000 ($\Delta_M \simeq 1{\cdot}5$ mm). The recording time was of the order of 30 minutes. In order to investigate the region 5000 to 6800 cm^{-1} with a resolving power of 1000, 6000 samples were taken from the interferogram, and the time required for the computation with the IBM 704 computer was of the order of 40 minutes. The apodization function was $T(\Delta) = [1-(\Delta/\Delta_M)^2]^2$. Figure 15 shows part of the spectrum which was obtained, in the region 6000 to 6500 cm^{-1}. Figure 16, which represents the same spectrum, in the region around 1 μ, shows strikingly the increase in resolution which results from increasing Δ_M; the two spectra, with resolving powers respectively of 300 and 1000, were obtained from the same interferogram, but using different lengths; Δ_M was made equal to 0·3 mm for the former and 1 mm for the latter.

In order to obviate the need of a beam-splitter in the far infra-red, Strong and his collaborators (Strong 1957, Strong and Vanasse 1956, Vanasse, Strong and Loewenstein 1959) have employed a laminated grating, with variable groove depth,

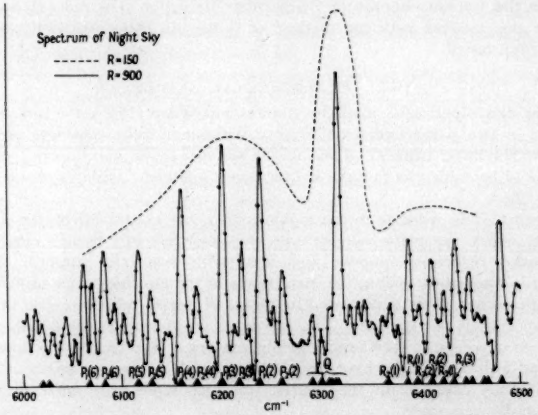


Figure 15. Example of spectrum obtained by Fourier transformation. The broken curve ($R = 150$) was obtained with a grating spectrometer.

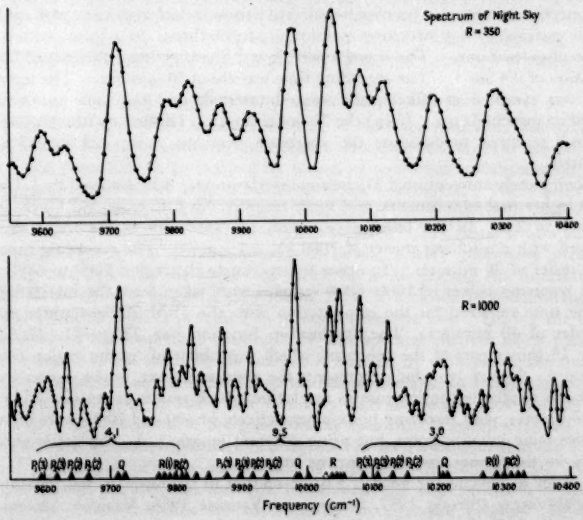


Figure 16. Example showing the increase in resolvance obtained by increasing Δ_M.

used with normal incidence. If d is the groove depth, and if only the light reflected back in the zero order of the grating is observed, the grating functions as a two-beam interferometer, giving a path difference of $2d\cos i$. The variable laminar grating consists of 24 movable plates which slide like pistons between 24 fixed plates. The modulator is situated similarly to the grating in an Ebert mounting; the slits must be narrow enough to eliminate the two first orders of the grating, and since the grating has only a very small number of elements, the slits have to be rather narrow, and consequently the solid angle available is much smaller than with a Michelson interferometer. In the most recent experiments the spectra were reconstituted by numerical calculation with the IBM 650; the spectra which have been published are absorption spectra of H_2O and H_2S in the region 200 μ with a theoretical limit of resolution of 0·2 cm^{-1}.

Any type of two-beam interferometer can be used, including those involving polarization. Mertz (1958) has constructed an instrument with a Soleil compensator between the analyser and the polarizer; the maximum change in the path difference is determined by the thickness of the compensator; it was 175 wavelengths at a wavelength of 4000 Å. The solid angle available is of the same order of magnitude as that of an equivalent Michelson interferometer. It is apparent that the apparatus is much simpler and more stable than the Michelson interferometer, and it possesses certain advantages in cases when only very low resolving power is required in the visible or in the near infra-red.

§ 18. Michelson's Method of Fringe Visibility

Let us consider again the general expression giving the flux, assuming now that the unit of luminance has been chosen so that

$$\int_0^\infty B(\sigma)\,d\sigma = 1\,;$$

the expression is then: $\phi(\Delta) = 1 + \int_0^\infty B(\sigma)\cos 2\pi\sigma\Delta\,d\sigma.$

In the general case, if $B(\sigma)$ extends over a wide domain, $\phi(\Delta)$ has a complicated form, and the interferogram shows no fringes; accordingly, it is impossible to see fringes when looking at the exit field of the interferometer.

However, if the spectrum $B(\sigma)$ extends over a narrow range around a central wave-number σ_0, $\phi(\Delta)$ becomes much simpler and it becomes possible to observe fringes. In effect, writing $\sigma = \sigma_0 + \sigma'$, and in terms of the displaced spectrum $B(\sigma')\,\phi(\Delta)$ becomes:

$$\phi(\Delta) = 1 + C\cos 2\pi\sigma_0\Delta - S\sin 2\pi\sigma_0\Delta$$

where $$C = \int_{-\infty}^{+\infty} B(\sigma')\cos 2\pi\sigma'\Delta\,d\sigma'$$

and $$S = \int_{-\infty}^{+\infty} B(\sigma')\sin 2\pi\sigma'\Delta\,d\sigma'.$$

Since $B(\sigma)$ is zero outside a narrow range around σ_0, the integrals C and S vary slowly compared with $\cos 2\pi\sigma_0\Delta$; thus, over a few periods of $\cos 2\pi\sigma_0\Delta$ the variation of $\phi(\Delta)$ is sinusoidal, and fringes are visible, with a 'visibility' determined by

the values of C and S. It is important to realize that this is true only because $B(\sigma)$ is restricted to a narrow range, and hence C and S vary slowly. If $B(\sigma)$ extends over a wide spectral domain, C and S may vary as a function of Δ almost as rapidly as $\cos 2\pi\sigma_0\Delta$, and nothing resembling fringes can be observed.

With a monochromatic line the flux would be $\phi(\Delta) = 1 + \cos 2\pi\sigma_0\Delta$, while in the present case it can be written as $\phi(\Delta) = 1 + V\cos(2\pi\sigma_0\Delta + \gamma)$ where $V^2 = C^2 + S^2$ and $\tan\gamma = S/C$. The contrast, or 'visibility' according to Michelson's definition

$$\Gamma = \frac{\Phi_{\max} - \Phi_{\min}}{\phi_{\max} + \phi_{\min}}$$

is then $\Gamma = V$, instead of $\Gamma = 1$ as in the monochromatic case.

The function $V(\Delta)$ is the envelope of the interferogram $\phi'(\Delta) = \phi(\Delta) - 1$ as defined in § 12.

From here, we must distinguish two cases, according to whether $B(\sigma')$ is symmetrical about σ_0 or not.

A.—*Symmetrical case.* Then $S = 0$ and

$$V = C = \int_{-\infty}^{+\infty} B(\sigma')\cos 2\pi\sigma'\Delta\, d\sigma'.$$

According to this equation, the visibility function $V(\Delta)$ is identical with the interferogram of a fictitious spectrum deduced from the true spectrum by subtraction of σ_0 from all the wave-numbers (transposed spectrum). Thus the visibility curve ignores the periodic structure of the interferogram (period $1/\sigma_0$) and retains only its interesting features. The spectrum $B(\sigma')$ may be deduced unambiguously from the visibility function, by the inverse Fourier transform of V, exactly as $B(\sigma)$ can be deduced by the inverse Fourier transform of $\phi'(\Delta)$. Since V is much simpler than ϕ' this operation is much easier. A known analytical form may even be recognized in $V(\Delta)$ and its transform $B(\sigma')$ immediately deduced without tedious numerical computation. For example, if V is an exponential function, the spectrum is a single line with a Lorentz line shape ; if V is a Gaussian function, the spectrum is a single line with a Doppler line shape ; if V is sinusoidal, the spectrum is a doublet ; if V is a constant plus a sine function, the spectrum is a triplet, and so on. In each of these very simple cases the parameters of the spectrum (line width, or line separation) may be deduced from the parameters of V. In addition any combination of these simple cases is reflected in V and can be easily recognized. Such a combination is a convolution, the transform of which is the product of the elementary transforms : terms of the product can be isolated and a number of features of the spectrum can be found (Terrien 1958, 1959).

The first application of this method was made by Michelson (1892) who measured the visibility visually : he was thus able, for instance, to show that the profile of the red cadmium line was nearly a pure Doppler profile and to measure its width. Recently, Terrien improved the method by measuring the visibility photoelectrically : he was able, for instance, to show that the profile of the orange line of krypton, emitted in certain conditions suitable for metrology, is a mixture of Doppler and Lorentz profiles the widths of which are in the ratio of about 10/1.

B.—*Non-symmetrical case.* In this case $S \neq 0$ and the knowledge of V alone is no longer sufficient to determine $B(\sigma')$. The spectrum $B(\sigma')$ can be split into a

symmetrical part B_s and an antisymmetrical part B_a related to C and S by the relations

$$B_s(\sigma') = \int C(\Delta) \cos 2\pi\sigma'\Delta \, d\Delta$$

and

$$B_a(\sigma') = \int S(\Delta) \sin 2\pi\sigma'\Delta \, d\Delta.$$

Hence in order to determine $B(\sigma')$, C and S must be known separately. Their determination is possible from the knowledge of $V(\Delta)$ and $\gamma(\Delta)$ through the relations $C = V\cos\gamma$ and $S = V\sin\gamma$.

The phase γ of the fringes must be measured by some method, for instance by comparing the fringes of the unknown spectrum with the fringes given by a monochromatic line, the symmetry of which has been separately tested.

Terrien proposes to call this method an ' extended Michelson's method ' ; it would permit the study of simple and narrow spectra more simply than the method involving the Fourier transform of the complete interferogram. In his famous work, Michelson discovered that such lines as the hydrogen or mercury lines have a structure. From the form of the visibility curve he inferred a (non-symmetrical) structure and adjusted the possible structure until it gave the observed visibility curve by Fourier transformation. But, as he did not use the information given by γ, the solution was not unique, and the structures found by Michelson were not correct.

Ingelstam (1954–1958) has proposed methods closely related to the method of fringe visibility : instead of a Michelson interferometer, he makes use of a device based on a phase contrast refractometer previously designed and used by him.

PART V. MISCELLANEOUS

§ 19. Michelson's Echelon and Harrison's Echelle

The well-known Michelson's echelon has been very little used during the last twenty years, nor have there been improvements in its conception or mode of use. As a matter of fact a reflecting echelon is nothing more than a grating used in a Littrow mounting with a good concentration of energy in one or two orders, a high blaze angle ϕ, and a very small number of grooves N. Since the resolving power of a grating in a Littrow mounting is $\mathscr{R} = 2W\sin\phi/\lambda$ the resolving power, and also the luminosity, of an echelon are the same as those of a grating having the same width W and angle ϕ, regardless of the interference order k ; and as a result of recent improvements in the manufacture of gratings (Harrison 1959) it is possible to obtain gratings approaching optical perfection with values of W and ϕ as high as those of a Michelson's echelon. On the other hand, the echelon suffers from difficulties due to the overlapping inherent in the use of high interference orders. These difficulties are exactly the same as in a Fabry–Perot of the same resolving power, since the number of steps N is about the same as the common value of the finesse N of a Fabry–Perot, but the echelon has not the advantage of the high luminosity of the Fabry–Perot. Before recent improvements in the gratings the only domain where the echelon was used with advantage was the ultra-violet region where no good coatings are yet available for the Fabry–Perot ; it is in this region

that it has been used by Hately and Littlefield (1958). Now it seems that the domain of use of the echelon is becoming more and more restricted.

The Harrison echelle (1949, 1952), which is half-way between an ordinary grating (that is, with low interference order) and an echelon, is valuable in the case where an extended spectrum is to be *photographed* at high resolvance in a single exposure, because it makes the best use of the two dimensions of the photographic plate. In an echelle spectrograph the dispersion of the echelle itself is crossed with the dispersion of an auxiliary disperser (prism or grating) so that the bandwidth of the auxiliary disperser is equal to the free spectral range $\Delta\sigma$ of the echelle. One thus obtains a two-dimensional array of spectral lines in which the dispersion and the resolution are very high parallel to the y axis (figure 17). However, the total

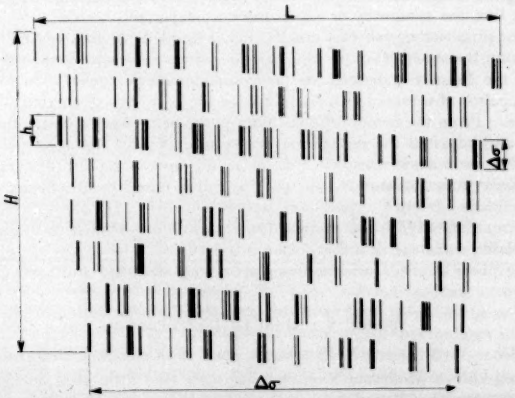


Figure 17. Two-dimensional array of lines obtained with an echelle crossed with an auxiliary disperser (L, H and h are linear dimensions on the plate ; $\Delta\sigma$ is the free spectral range of the echelle, equal to the bandwidth of the auxiliary disperser).

spectral range covered is very much extended, the length occupied by the spectrum on the plate being equal to the length L of each row (or cycle, containing one free spectral range) multiplied by the number of rows. The number of spectral elements in each cycle $\Delta\sigma$ being N (number of grooves), it is easily shown that the ratio H/L of the height H to the length L of the area occupied by the spectrum is :

$$H/L = b\beta k/N = b\beta\mathscr{R}/N^2$$

where b is the ratio of the total width $\sigma_2 - \sigma_1$ of the spectrum to the mean value σ, β is the ratio of the length h of each line to its width (assuming that the theoretical resolvance is approximately reached), k is the mean interference order, N is the number of grooves, $\mathscr{R}$ is the mean resolving power $\mathscr{R} = kN$.

Assuming definite values for $b, \beta, \mathscr{R}$, the value of N allowing a plate of a given H/L to be used may be determined according to this equation. If, for example,

$b = 1/2, \beta = 50$ (that means a line length of 1·5 mm if the grain size is 30 μ), $\mathscr{R} = 10^8$, $H/L = 3$ (which is a common value for a photographic plate), one obtains $N \simeq 3000$ (and, therefore, $k \simeq 300$). So, with the assumed data (rather arbitrary), $N = 3000$ is the value which allows the plate to be filled as efficiently as possible with the spectrum. As a matter of fact the echelles produced by Harrison have a number of grooves from 1000 to 6000. Thus the essential superiority of the echelle over the other types of grating is not in its resolving power or in its luminosity, but in the fact that it concentrates the most possible spectral information in a plate of common size. Of course this superiority holds only for the photographic use, and completely disappears in a spectrometer.

No further development will be given here on this type of grating, because, as stated in the introduction, gratings have been considered as falling outside of the scope of this report, although the echelle may be considered as being on the verge of interference spectroscopy.

§ 20. Lummer Plate

This device, which has played a brilliant role in the early history of high resolution spectroscopy, is nowadays completely abandoned. Like the Fabry–Perot étalon, the Lummer plate makes use of the process of luminance division for producing multiple beams. As a matter of fact there is no loss of luminance in the process involved, and the coefficient τ_A is equal to unity, whilst in the Fabry–Perot it rarely exceeds 0·7. But this slight superiority is off-set by many inferiorities. The symmetry of the device is more like that of a grating than that of a Fabry–Perot and the plate could not be used in a spectrometer to obtain the gain in luminosity of the Fabry–Perot. The law of dispersion is complicated, and, above all, the slightest defect in the figuring of the plate produces spurious effects and it is almost impossible to produce plates of the desirable optical quality.

§ 21. The Basic Reasons for the Superiority of Certain Devices

In the different parts of this report it has been noticed that the Fabry–Perot spectrometer, the SISAM, and the Fourier method using a Michelson interferometer allow a much higher solid angle than other devices for the same resolution. An obvious remark is that these three devices have a rotational symmetry (even the SISAM) and make use of the whole of a fringe (a ' closed fringe '), that is the maximum space where the phase difference has a given value. Of course this condition is a necessary one if the high solid angle has to be allowed, but at the same time it may be remarked that the three devices have in common the use of some semi-transparent beam splitting plate (luminance division). The question then arising is whether this condition is also a necessary one or is only a coincidence. A proof that the condition is indeed necessary has been given by Jacquinot (1958): the interference must take place between rays which have been derived from the same incident ray by luminance division by means of a semi-transparent plate or by polarization. A closed fringe embraces the preferred direction where all the interfering rays coincide with each other, i.e. the direction about which the variation of the phase difference of the interfering rays as a function of the angle is given by a second-order term. Obviously this preferred direction exists only if the division

of an incident ray into interfering rays is a luminance division. Thus, closed fringes exist only in the case of luminance division.

Hence, in order to admit a solid angle as large as $2\pi/\mathscr{R}$, both conditions are necessary : (*a*) luminance division by semi-transparent plates or by polarization ; (*b*) use of the whole of a closed fringe.

It is therefore hopeless to try to invent devices having rotational symmetry, like glass cones or circular grating but not fulfilling the condition (*a*).

The reason for the superiority in luminosity of the Fabry–Perot over the grating in the photographic use is not so obvious ; nevertheless it is related to the condition (*a*). Here, the high luminosity depends on the high dispersion as explained at the end of § 9 : the dispersion tends to infinity as the interfering rays tend to coincidence with each other in a Fabry–Perot used at normal incidence or in a grating used at grazing incidence. But when the incidence is such that the dispersion is infinite, the cross section of the emergent beam becomes zero with a grating, whereas it is a maximum with a Fabry–Perot.

Acknowledgments

The author would like to express his gratitude to Dr. R. Chabbal and Dr. P. Connes for their great help in the preparation of the manuscript and for many helpful discussions.

References

Adam, M. G., 1948, *Comm. Univ. Obs. Oxford*, No. 26 ; 1958, *Mon. Not. R. Astr. Soc.*, **118**, 106.

Baird, K. M., 1958, *J. Phys. Radium*, **19**, 384.

Barr, W. L., and Jenkins, F. A., 1956, *J. Opt. Soc. Amer.*, **46**, 141.

Baumeister, P. W., and Stone, J. M., 1956, *J. Opt. Soc. Amer.*, **46**, 228.

Blaise, J., 1957, *Thesis* (1958, *Ann. Phys., Paris*, **3**, 1019).

Blaise, J., Chabbal, R., and Jacquinot, P., 1954, *J. Phys. Radium*, **15**, 749.

Blaise, J., and Chantrel, H., 1953, *J. Phys. Radium*, **14**, 135.

Born, M., 1933, *Optik* (Berlin : Springer), p. 477.

Brochard, J., Chabbal, R., Chantrel, H., and Jacquinot, P., 1952, *J. Phys. Radium*, **13**, 433.

Brochard, J., Chantrel, H., and Jacquinot, P., 1958, *J. Phys. Radium*, **19**, 515.

Brody, J. K., Tomkins, F. S., and Fred, M., 1957, *Spectrochemica Acta*, **8**, 329.

Burridge, J. C., Kuhn, H., and Pery, A., 1953, *Proc. Phys. Soc.* B, **66**, 963.

Cartwright, C. H., and Turner, A. F., 1939, *Phys. Rev.*, **55**, 1128.

Chabbal, R., 1953, *J. Rech. C.N.R.S.*, **24**, 138 ; 1957, *Thesis* (1958, *Rev. Opt.*, **37**, 49, 336, 501).

Chabbal, R., and Jacquinot, P., 1954, Communication aux Congrès de Lund et Parme (1955, *Nuovo Cim.*, **2**, 661).

Chabbal, R., and Noorman, P., 1958, *J. Phys. Radium*, **19**, 371.

Chabbal, R., and Soulet, M., 1958, *J. Phys. Radium*, **19**, 274.

Chantrel, H., 1958, *J. Phys. Radium*, **19**, 366 ; 1959, *Thesis* (1959, *Ann. Phys., Paris*, **4**, 965).

Connes, J., 1958, *J. Phys. Radium*, **19**, 197.

Connes, J., and Gush, H., 1959, Symposium on Interferometry, Teddington, G.B.

Connes, P., 1956, *Rev. Opt.*, **35**, 37 ; 1958, *J. Phys. Radium*, **19**, 262.

Courtès, G., 1951, *C.R. Acad. Sci., Paris*, **231**, 1283 ; 1958, *J. Phys. Radium*, **19**, 342.

Culshaw, W., Richardson, J. M., and Kerns, D. M., 1959, Symposium on Interferometry, Teddington, G.B.

Dufour, C., 1949, *Thesis* (1951, *Ann. Phys., Paris*, **6**, 5).

DUFOUR, C., and JACQUINOT, P., 1948, *J. Rech.*, **6**, 1.
DUFOUR, C., and PICCA, R., 1945, *Rev. Opt.*, **24**, 19.
DUPEYRAT, R., 1957, *Thesis* (1958, *Ann. Phys., Paris*, **3**, 377); 1958, *J. Phys. Radium*, **19**, 290, 351.
FELLGETT, P., 1952, Symposium Ohio State Univ.; 1958, *J. Phys. Radium*, **19**, 187, 237.
GEAKE, J. E., and WILCOCK, W. L., 1956, *Mon. Not. R. Astr. Soc.*, **111**, 561.
GEBBIE, H. A., 1958, *J. Phys. Radium*, **19**, 230; 1959, Symposium on Interferometry, Teddington, G.B.
GIACOMO, P., 1955, *Thesis* (1956, *Rev. Opt.*, **35**, 317, 442); 1958, *J. Phys. Radium*, **19**, 307.
GIACOMO, P., and JACQUINOT, P., 1952, *J. Phys. Radium*, **13**, 59A.
GOBERT, J., 1958, *J. Phys. Radium*, **19**, 278.
GREENLER, R. G., 1955, *J. Opt. Soc. Amer.*, **45**, 788; 1958, *J. Phys. Radium*, **19**, 375.
GUILINO, G., and MAIER LEIBNITZ, H., 1958, *J. Phys. Radium*, **19**, 270.
HARRISON, G. R., 1952, *J. Opt. Soc. Amer.*, **39**, 522.
HARRISON, G. R., ARCHER, J. E., and CAMUS, J., 1952, *J. Opt. Soc. Amer.*, **42**, 706.
HARRISON, G. R., STURGIS, N., DAVIS, S. P., and YAMATA, Y., 1959, *J. Opt. Soc. Amer.*, **49**, 205.
HATELY, G. F., and LITTLEFIELD, T. A., 1958, *J. Opt. Soc. Amer.*, **48**, 851.
HIRSCHBERG, J. G., 1958, *J. Phys. Radium*, **19**, 256.
HIRSCHBERG, J. G., KADESCH, R., and MACK, J. E., 1958, *J. Phys. Radium*, **19**, 338.
HUMPHREYS, C. J., and PAUL, E., 1955, Pittsburgh Meeting; 1958, *J. Phys. Radium*, **19**, 424.
INGELSTAM, E., 1954, *Optica Acta*, **1**, 73; 1958, *J. Phys. Radium*, **19**, 202, 209.
JACKSON, D. A., 1956, *Phys. Rev.*, **103**, 1738; 1957, *Proc. Roy. Soc.* A, **241**, 283; 1958, *J. Phys. Radium*, **19**, 379.
JACKSON, D. A., and KUHN, H. G., 1938, *Proc. Roy. Soc.* A, **167**, 205; 1939, *Ibid.*, **173**, 278.
JACQUINOT, P., 1950, *Proc. Phys. Soc.* B, **63**, 969; 1954 a, XVIIème Congrès du G.A.M.S., Paris; 1954 b, *J. Opt. Soc. Amer.*, **44**, 761; 1958, *J. Phys. Radium*, **19**, 223; 1960, *Optica Acta*, in the press.
JACQUINOT, P., and DUFOUR, C., 1948, *J. Rech. C.N.R.S.*, **6**, 1.
JAFFÉ, J. K., 1958, *J. Phys. Radium*, **19**, 241, 273.
JAFFÉ, J. K., RANK, D. H., and WIGGINS, T. A., 1955, *J. Phys. Radium*, **45**, 636.
JARRETT, A. H., and VON KLUBER, H., 1955, *Mon. Not. R. Astr. Soc.*, **115**, 343.
JENKINS, F. A., 1948, *J. Phys. Radium*, **19**, 301.
JENKINS, F. A., and BARR, W. L., 1956, *J. Opt. Soc. Amer.*, **46**, 141.
KALITEEVSKY, N. I., and TCHAÏKA, M. P., 1956, *Vest Lenin Ouniver.*, **4**.
KLINE, J. V., 1956, *Thesis*, Purdue University.
KOROLEV, F. A., 1953, *High Resolution Spectroscopy—GITTL.*
KUHN, H. G., 1951, *Reports on Progress in Physics*, **14**, 64.
KUHN, H. G., and LUCAS-TOOTH, H. J., 1958, *J. Phys. Radium*, **19**, 293.
KUHN, H. G., and RAMSDEN, S. A., 1958, *J. Phys. Radium*, **19**, 383.
KUHN, H. G., and WILSON, B. A., 1950, *Proc. Phys. Soc.* B, **63**, 745.
LENNUIER, R., 1958, *J. Phys. Radium*, **19**, 319.
LENNUIER, R., and COJAN, J. L., 1952, *C.R. Acad. Sci., Paris*, **235**, 1634.
LITTLEFIELD, T. A., and BLAKIE, J., 1955, *J. Opt. Soc. Amer.*, **45**, 658.
MERTZ, L., 1958, *J. Phys. Radium*, **19**, 233; 1959, Conference I.C.O., Stockholm.
MICHELSON, A., 1892, *Phil. Mag.*, **34**, 280.
MINKOWSKI, R., and BRUCK, H., 1935, *Z. Phys.*, **27**, 95.
PENSELIN, S., and STEUDEL, A., 1955, *Z. Phys.*, **142**, 21.
RANK, D. H., and SHEARER, J. N., 1956, *J. Opt. Soc. Amer.*, **46**, 463.
RAYLEIGH, LORD, 1892, *Phil. Mag.*, **34**, 407.
RING, J., BEER, R., and HEWISON, V., 1958, *J. Phys. Radium*, **19**, 325.
RING, J., and LISSBERGER, P. H., 1955, *Optica Acta*, **2**, 43.
RING, J., and WILCOCK, W. L., 1954, *Nature, Lond.*, **173**, 994.
RING, J., and WOOLF, N. J., 1958, *J. Phys. Radium*, **19**, 354.
RUBENS, H., and WOOD, R. W., 1911, *Phil. Mag.*, **21**, 249.
STEUDFL, A., 1958, *J. Phys. Radium*, **19**, 312.
STRONG, J., 1957, *J. Opt. Soc. Amer.*, **47**, 354.

STRONG, J., and VANASSE, G., 1956, *Nature, Lond.*, **178**, 432; 1958, *J. Phys. Radium*, **19**, 192.

TCHAÏKA, M. P., 1957, *Optika i Spectroscopia*, **3**, 372.

TERRIEN, J., 1958, *J. Phys. Radium*, **19**, 390; 1959, Symposium on Interferometry, Teddington, G.B.

TOLANSKY, S., and BRADLEY, D. J., 1959, Symposium on Interferometry, Teddington, G.B.

TREANOR, P. J., 1948, *Nature, Lond.*, **162**, 885; 1957, *Mon. Not. R. Astr. Soc.*, **117**, 22.

VAN DER SLUIS, K. L., and MACNALLY, 1956, *J. Opt. Soc. Amer.*, **46**, 39.

VANASSE, G., STRONG, J., and LOEWENSTEIN, E., 1959, *J. Opt. Soc. Amer.*, **49**, 309.

WARK, D. Q., and STONE, J. M., 1955, *Nature, Lond.*, **175**, 254.

PLANETARY NEBULAE

By M. J. SEATON

Department of Physics, University College London

CONTENTS

	PAGE
§ 1. Introduction	314
§ 2. Hydrogen and helium spectra	316
2.1. Optical depths	316
2.2. Hydrogenic recombination spectra	317
2.3. Allowance for orbital degeneracy	320
2.4. The He I recombination spectrum	321
2.5. Summary of calculations	321
2.6. Comparisons with observations	322
2.7. Determination of He^{+2} and He^{+} abundances	323
2.8. Continuous spectra of planetaries	325
2.9. The absolute brightness	328
2.10. Collisional excitation of hydrogen	328
§ 3. Transfer problems	329
3.1. The Lyman continuum	329
3.2. The He I Lyα line	329
3.3. The He II Lyα line	330
§ 4. The forbidden lines	331
4.1. Excitation mechanisms	331
4.2. Equilibrium equations and intensity ratios	332
4.3. Determination of temperatures and densities	333
§ 5. Fluorescent mechanisms	334
5.1. The O III lines	335
5.2. The N III lines	336
5.3. Discussion of the fluorescent efficiencies	336
§ 6. Chemical compositions	336
6.1. Ionic abundances from recombination lines	337
6.2. Ionic abundances from forbidden lines	338
6.3. Determination of total abundances	338
6.4. Results for NGC 7027	338
§ 7. Sizes of ionized regions	340
7.1. Optical depths for H absorption	340
7.2. The transition region	340
7.3. Radii of ionized regions	341
7.4. Determination of He/H abundance ratios	343
§ 8. Star temperature	344
8.1. Brightness of the star	344
8.2. Numbers of ionizing quanta	344
8.3. The thermal balance	345
8.4. Summary of theory relations	346
8.5. NGC 6572 and 7027	347
8.6. Discussion of the stellar ultra-violet	349
§ 9. The equation of state	349
References	352

Abstract. Planetary nebulae are gas clouds surrounding certain hot stars. The observational data are interpreted in terms of the processes taking place in a low density ionized gas exposed to dilute ultra-violet radiation. The aim is to obtain information about the density, kinetic temperature and chemical composition of the nebular material and about the nature of the ultra-violet radiation field.

§ 1. Introduction

Planetary nebulae are clouds of ionized gas surrounding certain hot stars. They have dimensions of order 10^{17} cm or 10^{6} solar radii. Seen through a telescope they appear as pale green discs resembling the planets Uranus and Neptune. The light emitted by planetaries is due to absorption of the ultra-violet radiation of the central star, this radiation being diluted by a factor $W = R_*^2/4R^2$ where R is the distance from the star and R_* the star radius. Since W is very small ($\sim 10^{-13}$) conditions are very different from those holding in thermodynamic equilibrium. The electron densities N_e are of order $10^4\,\mathrm{cm}^{-3}$ and the kinetic temperatures T_e of order 10^4 deg K. Hydrogen is the most abundant element. There is about one atom of helium for every five of hydrogen. All other elements are much less abundant; thus there is only one atom of oxygen to every 1000 atoms of hydrogen.

The spectrum of a planetary nebula contains a large number of emission lines and a background continuum. Our aim is to interpret this spectrum in terms of the processes taking place in a low density gas exposed to dilute ultra-violet radiation and to obtain information about the temperature, density and chemical composition of the nebula and the nature of the ultra-violet radiation of the central star.

The physical processes in planetaries are similar to those in the diffuse emission nebulae, of which the Orion nebula is the best known example. The diffuse nebulae have densities ranging between 10^4 electrons per cm^3 and one electron per cm^3.

The distinguishing feature of the planetary nebulae is the close connection between nebula and star. It is probable that the nebular gas has been ejected from the star, the mass of the nebula being only a fraction of the mass of the star. Spectral lines for light coming from the centre of the disc of a planetary are often seen to be split when examined at high dispersion (Wilson 1958). This shows that planetaries are expanding, typical velocities being of order 20 km sec^{-1} (the nebulae are transparent to the observed radiations and the two components of lines are due to light from the far side and light from the near side). The large diffuse nebulae differ from the planetaries in that they are usually associated with a number of hot stars and have masses many times greater than the mass of a single star. Here the stars may have condensed out of the gaseous medium. Emission nebulae also occur in extra-galactic objects, in some cases the entire spectrum being dominated by characteristic nebular emission lines.

The primary physical process in a planetary is absorption of stellar ultra-violet in atomic photo-ionization:

$$\mathrm{H} + h\nu \rightarrow \mathrm{H}^+ + \mathrm{e}. \qquad \ldots\ldots(1)$$

This usually takes place from ground state atoms. Photo-ionization provides the source of random kinetic energy in the nebula. Loss of kinetic energy results from inelastic collisions followed by radiation. Redistribution of kinetic energy in elastic collisions between charged particles is very effective in maintaining a Maxwell velocity distribution (Bohm and Aller 1947, Spitzer 1956) and the rates of energy gain and energy loss therefore determine a kinetic temperature T_e.

The most important inelastic collisions are those producing excitation of low-lying metastable states of various atoms and ions. The strongest lines, responsible

for the pale green appearance, are emitted following the excitation process

$$O^{+2}(2p^2\,{}^3P) + e \rightarrow O^{+2}(2p^2\,{}^1D) + e. \qquad \ldots\ldots(1.2)$$

These lines (known as N_1, λ 5007, $^1D \rightarrow {}^3P_2$ and N_2, λ 4959, $^1D \rightarrow {}^3P_1$) were observed long before the identification was established (Bowen 1928) and were first ascribed to 'nebulium'. The upper state, O^{+2} $2p^2$ ^1D, has a radiative lifetime of 36 sec (Garstang 1951) compared with about 10^{-8} sec for typical non-metastable states. In addition to the† [O III] lines many other forbidden lines are observed ([O I], [O II], [N II], [S II], ...). Although the low density in planetaries favours the emission of forbidden lines, the probabilities of collisional deactivation are by no means always negligible compared with the probabilities of forbidden line emission. We can often find two different forbidden lines emitted by the same ion. The intensity of these lines will depend on both T_e and N_e and by studying a number of such ratios T_e and N_e may be deduced (§ 4). The energies ΔE required to excite the forbidden lines are of order 2 ev and it is energy differences of this order which control the kinetic temperature, which adjusts itself to be such that $kT \simeq \Delta E$ (§ 9).

The forbidden lines are strong because ΔE is small and despite the fact that the emitting ions have small abundances. One also observes permitted lines due to recombination processes,

$$H^+ + e \rightarrow H(nl) + h\nu. \qquad \ldots\ldots(1.3)$$

These are strong only when the recombining ion (H^+, He^+ or He^{+2}) has a large abundance, but many weak recombination lines are observed for other ions (O^{+2}, O^{+3}, O^{+4}, O^{+5}, C^{+2}, C^{+3}, ...).

A third, and highly selective, excitation process is by fluorescent mechanisms (Bowen 1934). It happens that the Lyα line of He II has a wavelength very close to a resonance line of O III and that a line of O III which is emitted following fluorescent excitation has a wavelength very close to a resonance line of N III, so that fluorescent lines of N III are also produced.

The observed light of planetaries is produced in regions of ionized gas. We must consider what determines the sizes of such regions. One possibility is that the boundary of the ionized region is the boundary of the gas cloud. A second possibility is that the gas cloud is larger than the ionized region, the extent of the latter being determined by the number of available ionizing quanta. It is important to realize that the second possibility would give a sharp outer boundary, beyond which the ionizing radiation could not penetrate. The explanation is that as soon as the ionizing radiation becomes appreciably attenuated by absorption the number of neutral hydrogen atoms will increase and this in turn will give an increased absorption coefficient. This effect causes the attenuation by absorption to increase rapidly once it starts to be appreciable.

Problems of planetary nebulae are discussed in books by Vorontsov-Velyaminov (1948), by Wurm (1951) and by Aller (1956). A series of eighteen papers entitled "Physical Processes in Gaseous Nebulae" appeared in the *Astrophysical Journal* between 1937 and 1945; references are given by Aller and Menzel (1945). A book by Dufay (1957) deals with problems of planetary and diffuse nebulae and

† We use the Roman numeral notation (O I, O II, O III, ...) only when referring to the spectrum (of O^0, O^+, O^{+2}, ...). The square brackets, as in [O III], are used for forbidden lines.

with other problems of interstellar matter, such as the neutral gas and the interstellar dust grains. The neutral gas is observed by absorption lines (Na I, Ca II, ...) and by the 21 cm emission line. The interstellar dust clouds may be observed by reflection of starlight or by absorption. The dust absorption increases with increasing frequency and therefore gives a reddening effect which must be taken into account when discussing the light received from distant objects.

§ 2. Hydrogen and Helium Spectra

2.1. *Optical Depths*

The cross section for photo-ionization from level n of a hydrogenic atom with nuclear charge Z is

$$a_n(Z,\epsilon) = \left(\frac{2^6\,\alpha\pi a_0^2}{3\sqrt{3}}\right)\frac{n}{Z^2}(1+n^2\epsilon)^{-3}g_{II}(n,\epsilon) \qquad \ldots\ldots(2.1)$$

where

$$\left(\frac{2^6\,\alpha\pi a_0^2}{3\sqrt{3}}\right) = 7{\cdot}907\times10^{-18}\,\mathrm{cm}^2 \qquad \ldots\ldots(2.2)$$

(α being the fine structure constant and a_0 the Bohr radius). The photon energy is

$$h\nu = Z^2 I_H\left(\frac{1}{n^2}+\epsilon\right), \qquad \ldots\ldots(2.3)$$

I_H (= 13·60 ev) being the threshold ionization energy of hydrogen. The ejected electron energy is $Z^2\epsilon$ in units of 13·60 ev. In (2.1), g_{II} is the Kramers–Gaunt factor of order unity ; it may be represented by the asymptotic expansion (Menzel and Pekeris 1935, Burgess 1958):

$$g_{II}(n,\epsilon) = 1+0{\cdot}1728\,n^{-2/3}(u+1)^{-2/3}(u-1) - 0{\cdot}0496\,n^{-4/3}(u+1)^{-4/3}(u^2+\tfrac{4}{3}u+1)+\ldots \qquad \ldots\ldots(2.4)$$

where $u = n^2\epsilon$. Substitution of $n = 1$, $Z = 1$, and $\epsilon = 0$ gives† $a_1(1,0) = 6{\cdot}15\times10^{-18}$ cm² for the threshold of the hydrogen Lyman continuum. The corresponding optical depth is $\tau_1 = N_1 a_1 R$ where N_1 is the density of ground state H atoms and R the nebular radius.

If a nebula is optically thick in the Lyman continuum ($\tau_1 > 1$) it will be very thick in the Lyman lines. For a transition $n\to n'$ the absorption cross section, assuming thermal Doppler broadening, is

$$\kappa_\nu(n,n') = \kappa_0(n,n')\exp\{-b(\nu-\nu_{nn'})^2\} \qquad \ldots\ldots(2.5)$$

where

$$\kappa_0(n,n') = \frac{\pi e^2}{m\nu_{nn'}}\left(\frac{M}{2\pi kT}\right)^{1/2}f(n,n') \qquad \ldots\ldots(2.6)$$

$$b = \left(\frac{c}{\nu_{nn'}}\right)^2\frac{M}{2kT}, \qquad \ldots\ldots(2.7)$$

m being the electron mass, M the ion mass and $f(n,n')$ the oscillator strength. Hydrogenic f-values are given by

$$f(n,n') = \frac{2^5}{3\pi\sqrt{3}}\frac{nn'}{n'^2-n^2}g_1(n,n'). \qquad \ldots\ldots(2.8)$$

† The exact H 1s threshold cross section is $6{\cdot}30\times10^{-18}$ cm², the use of (2.4) giving an error of 2·5%. For conditions in planetaries, (2.4) will usually give rise to errors less than this.

The asymptotic expansion for g_1 is obtained from (2.4) on putting $u = -(n/n')^2$. Tables of g_1 are given by Baker and Menzel (1938) and a very complete tabulation of hydrogenic transition probability data is given by Green, Rush and Chandler (1957). Table 1 gives values of $\kappa_0(1, n')/a_1(1, 0)$ where $\kappa_0(1, n)$, the absorption cross section at the line centre, is calculated for $T = 10^4$ deg K.

Table 1. Ratios of Lyman line to Lyman Continuum Absorption Coefficients ($\kappa_0(1, n')$ for line centre at 10^4 deg K)

n'	2	3	4	5	6	7	8	9	10
$\kappa_0(1, n')/a_1(1, 0)$	9350	1500	520	240	135	82	54	38	27

We consider two extreme cases (Baker and Menzel 1938). Case A applies for a nebula which is optically thin for Lyman radiation and Case B for one which is optically thick. We could, of course, also consider intermediate cases of nebulae thin for the Lyman continuum but thick for the lower Lyman lines. A further intermediate possibility would be a nebula with velocity gradients sufficiently large for the effective optical depths to be small for the lines yet still large for the continuum.

2.2. *Hydrogenic Recombination Spectra*

We first consider Case A. The number of recombinations on level n per cm^3 per sec is $N_+ N_e \alpha_n(T_e)$ where N_+ is the number of recombining ions per cm^3. The recombination coefficient $\alpha_n(T_e)$ is related to the photo-ionization cross section $a_n(h\nu)$ by

$$\alpha_n(T) = \left(\frac{2}{\pi}\right)^{1/2} \frac{\omega_n}{c^2 (mkT)^{3/2}} \exp(I_n/kT) \int_{I_n}^{\infty} (h\nu)^2 \exp(-h\nu/kT)\, a_n(h\nu)\, d(h\nu) \qquad \ldots\ldots(2.9)$$

where ω_n is the statistical weight and I_n the threshold ionization energy of level n. For hydrogenic ions $\omega_n = 2n^2$ and $I_n = Z^2 I_H/n^2$. Using (2.1) one obtains

$$\alpha_n(T) = \mathscr{D} Z \lambda^{1/2} \left(\frac{x_n S_n(\lambda)}{n}\right) \qquad \ldots\ldots(2.10)$$

where

$$\mathscr{D} = \frac{2^6}{3}\left(\frac{\pi}{3}\right)^{1/2} \alpha^4 c a_0^2 = 5.197 \times 10^{-14}\ \text{cm}^3\,\text{sec}^{-1}, \qquad \ldots\ldots(2.11)$$

$$\lambda = \left(\frac{Z^2 I_H}{kT}\right) = 157890 \left(\frac{Z^2}{T}\right) \quad (T \text{ in } ^\circ\text{K}), \qquad \ldots\ldots(2.12)$$

$$x_n = (I_n/kT) = \lambda/n^2 \qquad \ldots\ldots(2.13)$$

and

$$S_n(\lambda) = \int_0^\infty \frac{g_{II}(n, \epsilon) \exp(-x_n u)}{1+u}\, du. \qquad \ldots\ldots(2.14)$$

Tables for the calculation of $S_n(\lambda)$ are given by Seaton (1959 a). Table 2 gives some values of recombination coefficients.†

† We use the notation $\alpha_n(X^{+m})$ for recombination to the level n of X^{+m}.

Table 2. Recombination Coefficients α_1 and $\alpha_B = \sum_{n=2}^{\infty} \alpha_n$ for hydrogenic ions (α_n in cm^3 sec^{-1})

T_e (°K)	1×10^4	2×10^4
$10^{14}\,\alpha_1(H^0)$	15·5	10·5
$10^{14}\,\alpha_B(H^0)$	25·8	14·3
$10^{14}\,\alpha_1(He^+)$	63·6	44·5
$10^{14}\,\alpha_B(He^+)$	154·0	90·1

Any level n is populated by direct capture on n and by capture on higher levels n'' followed by cascade to n. Let $C_{n''n}$ be the probability that capture on n'' is followed by a transition to n when all cascade routes are taken into account.† The total rate of population of n is

$$N_+ N_e \sum_{n'' \geq n} \alpha_{n''} C_{n''n} \qquad \ldots\ldots(2.15)$$

where $C_{nn} = 1$. Let $A_{nn'}$ be the probability per second of a radiative transition from n to n' and define

$$A_n = \sum_{n'=1}^{n-1} A_{nn'} \qquad \text{(Case A)} \qquad \ldots\ldots(2.16)$$

$$P_{nn'} = A_{nn'}/A_n. \qquad \ldots\ldots(2.17)$$

Then

$$C_{n''n} = \sum_{m=n}^{n''-1} P_{n''m} C_{mn} \qquad \ldots\ldots(2.18)$$

and $C_{n''n}$ may therefore be calculated by a systematic procedure once the $P_{nn'}$ are known. Some tables of $C_{n''n}$ are given by Seaton (1959 b). Let $\mathcal{Q}_{nn'}$ be the number of quanta emitted in the $n \to n'$ line per cm^3 per sec; then $\mathcal{Q}_{nn'}$ is equal to $P_{nn'}$ times the rate of population of n. Let us put

$$\mathcal{Q}_{nn'} = N_+ N_e \alpha_{nn'} \qquad \ldots\ldots(2.19)$$

where $\alpha_{nn'}$ is the *effective recombination coefficient* for the $n \to n'$ line. We then have

$$\alpha_{nn'} = P_{nn'} \sum_{n'' \geq n} \alpha_{n''} C_{n''n}. \qquad \ldots\ldots(2.20)$$

Results of the calculations are usually expressed in terms of the parameters b_n which measure the departures from thermodynamic equilibrium. These are defined by

$$N_n = N_+ N_e \left(\frac{\boldsymbol{h}^2}{2\pi m \boldsymbol{k} T}\right)^{3/2} \frac{\omega_n}{2} b_n \exp(x_n), \qquad \ldots\ldots(2.21)$$

where N_n is the number of atoms per cm^3 in level n. With $b_n = 1$ equation (2.21) is the Saha ionization equation for thermodynamic equilibrium. Comparing (2.21) with

$$N_n A_n = N_+ N_e \sum_{n'' \geq n} \alpha_{n''} C_{n''n} \qquad \ldots\ldots(2.22)$$

we obtain

$$b_n \exp(x_n) = \left(\frac{2\pi m \boldsymbol{k} T}{\boldsymbol{h}^2}\right)^{3/2} \frac{1}{n^2 A_n} \sum_{n'' \geq n} \alpha_{n''} C_{n''n}. \qquad \ldots\ldots(2.23)$$

† This was first used by Plaskett (1928).

Values of b_n obtained on taking an infinite number of levels n'' into account have been calculated by Baker and Menzel (1938) and somewhat more accurately by Seaton (1959 b). In terms of the $b_n \exp(x_n)$ and the Kramers–Gaunt factors $g_1(n, n')$ (Baker and Menzel 1938) the effective recombination coefficients are given by

$$\alpha_{nn'} = 2\mathscr{D} \frac{g_1(n, n')}{nn'(n^2 - n'^2)} Z\lambda^{3/2} b_n \exp(x_n). \qquad \ldots\ldots(2.24)$$

For Case A the total recombination rate is

$$\alpha_A = \sum_{n=1}^{\infty} \alpha_n. \qquad \ldots\ldots(2.25)$$

Since all recombinations eventually give a transition to the $n = 1$ ground state, the total emission rate $\mathscr{Q}$(Ly) for all Lyman quanta is equal to the total recombination rate, $N_+ N_e \alpha_A$.

We now consider Case B. When the nebula is thick in the Lyman lines absorptions from the ground state also contribute to the rate of population of level n. We take the optical thickness to be so large that the rate of absorption in each volume element is equal to the rate of transitions to the ground state so that, when writing the equilibrium equations, we may neglect both absorptions from the ground state and transitions to the ground state. The only change in the formulae already given is that A_n is now defined by

$$A_n = \sum_{n'=2}^{n-1} A_{nn'} \qquad \text{(Case B)} \quad \ldots\ldots(2.26)$$

We consider, in effect, a hydrogen atom without a 1s ground state.

For Case B all recombinations on $n' \geqslant 2$ eventually give a transition to $n = 2$. The emission rate $\mathscr{Q}$(Ba) for all Balmer quanta is therefore

$$\mathscr{Q}(\text{Ba}) = N_+ N_e \alpha_B \qquad \ldots\ldots(2.27)$$

where

$$\alpha_B = \sum_{n=2}^{\infty} \alpha_n. \qquad \ldots\ldots(2.28)$$

Since population of $n = 2$ is followed by emission of a Lyα quantum,

$$\mathscr{Q}(\text{Ly}\alpha) = \mathscr{Q}(\text{Ba}).$$

In Case B Lyα quanta are the only Lyman quanta emitted by the nebula, all other Lyman quanta being degraded to Lyα by scattering processes (Zanstra 1927). Consider the fate of a Lyman continuum quantum. Since the optical depth is large this will ionize a hydrogen atom before it can escape from the nebula. There is then a probability of about 0·4 of recombination to the ground state and of 0·6 of recombination to an excited state. If recombination is to the ground state another Lyman continuum quantum is emitted and this ionizes another hydrogen atom. Eventually recombination occurs to an excited state and a continuum quantum is emitted which has an energy too small for ionization of a ground state atom and which therefore escapes from the nebula. But population of an excited state eventually leads to population of $n = 2$ and emission of a Lyα quantum. The Lyman continuum quantum has therefore been degraded to a Lyα quantum.

The possibility has been considered (Greenstein 1946, Seaton 1954 c, Pottasch 1958) that absorption of Lyα quanta in Case B might result in the population of $n = 2$ becoming sufficiently large for a nebula to become optically thick in the Balmer series. In the extreme case of a nebula optically thick for all Balmer transitions, all Balmer quanta will be degraded to Hα which will be the only Balmer quanta observed. Consider a less extreme case of a nebula with optical depth unity for Hβ. There will then be appreciable degradation of Hβ to Hα but since the absorption coefficients decrease rapidly with n all other Balmer lines will be little affected. The only change in the observed spectrum will be a decrease in Hβ and an increase in Hα. This trend is not observed in planetaries and we may conclude that the Hβ optical depth is less than one.

2.3. *Allowance for Orbital Degeneracy*

The theory considered so far is incomplete in that we have neglected the orbital degeneracy of the H atom. For thermodynamic equilibrium we have

$$N_{nl} = \frac{2l+1}{n^2} N_n. \qquad \ldots\ldots(2.29)$$

This relation has been tacitly assumed in the above discussion: thus the $A_{nn'}$ are defined by

$$A_{nn'} = \frac{1}{n^2} \sum_{l'} (2l+1) A_{nl,\, n'l'} \qquad \ldots\ldots(2.30)$$

where the upper states are weighted by the factor $2l+1$. Since conditions in a nebula will be far from thermodynamic equilibrium, equation (2.29) will not be valid. Approximate calculations taking the individual nl states into account have been made by Burgess (1958) and by Searle (1958). Exact calculations have been made by Burgess and Pengelly (unpublished). Table 3 gives some values of the

Table 3. Values of b_{nl} for $T_e = 1 \times 10^4$ deg K

	Case A				
n \ l	0	1	2	3	4
2	—	0·00386			
3	0·647	0·0270	0·0674		
4	0·882	0·0536	0·135	0·172	
5	1·06	0·0745	0·188	0·276	0·212
	Case B				
n \ l	0	1	2	3	4
3	1·03	0·234	0·0680		
4	1·16	0·338	0·137	0·172	
5	1·27	0·411	0·191	0·276	0·212

factors b_{nl}. These would be independent of l if (2.29) were true; this is seen to be far from being the case. So far as the observed line intensities are concerned it turns out that calculations assuming (2.29) give quite good results for Case B but less good results for Case A. In the exact calculations there are smaller differences between Cases A and B. The reason for this is readily understood.

Transitions to the ground 1s state take place only from the np states and, when (2.29) is assumed, the calculations over-estimate the importance of such transitions for Case A.

In § 2.8 we shall require the total effective recombination coefficient, $\alpha_{(2s)}$, for populating the 2s state. This is conveniently written in the form

$$\alpha_{(2s)} = X\alpha_B \qquad \ldots\ldots(2.31)$$

where α_B is defined by (2.28).

Values of X given in table 4 have been calculated by Burgess (1958) from the radiative theory. It will be shown later (§§ 2.8 and 4.3) that collisional processes may produce different effective values of X.

Table 4

	$10^{-4}\,T_e$	X		$10^{-4}\,T_e$	X
Case A	1	0·13	Case B	1	0·34
	2	0·15		2	0·37

For Case B the total emission rate for Balmer quanta is still given by $\mathcal{Q}(\mathrm{Ba}) = N_+ N_e \alpha_B$ when allowance is made for orbital degeneracy and the rate of $2 \rightarrow 1$ transitions is still equal to $\mathcal{Q}(\mathrm{Ba})$. We must, however, now distinguish between $2p \rightarrow 1s$ and $2s \rightarrow 1s$ transitions. The $2p \rightarrow 1s$ transitions give Lyα quanta with an emission rate $\mathcal{Q}(\mathrm{Ly}\alpha) = (1-X) N_+ N_e \alpha_B$. The $2s \rightarrow 1s$ transitions occur with emission of two photons (§ 2.8), the two-photon emission rate being $\mathcal{Q}(2s \rightarrow 1s) = X N_+ N_e \alpha_B$.

2.4. *The He I Recombination Spectrum*

The observed He I lines belong to $ns \rightarrow 2p$ and $nd \rightarrow 2p$ series. For $nd \rightarrow 2p$ we may use hydrogenic data calculated with allowance for l-degeneracy. Putting in statistical weight factors we have

$$\left.\begin{aligned} \alpha_{n\,^1\mathrm{D},\,2\,^1\mathrm{P}}(\mathrm{He}^0) &= \tfrac{1}{4}\alpha_{nd,\,2p}(\mathrm{H}^0), \\ \alpha_{n\,^3\mathrm{D},\,2\,^3\mathrm{P}}(\mathrm{He}^0) &= \tfrac{3}{4}\alpha_{nd,\,2p}(\mathrm{H}^0). \end{aligned}\right\} \qquad \ldots\ldots(2.32)$$

The He I $nd \rightarrow 2p$ intensities are insensitive to whether we assume Case A or Case B and the relative intensities are insensitive to T_e. Our results for He I are similar to those of Mathis (1957 a, b, c), but are probably a little more accurate.

2.5. *Summary of Calculations*

Table 5 gives calculated H I intensities $I(n, 2)$ relative to $I(4, 2) = 100$ where (4, 2) is the Hβ line, λ4861. The main calculations are made by the method of § 2.2; the figures in brackets are calculated as in § 2.3 allowing for l-degeneracy. Table 6 gives He II intensities $I(n, 4)$ and $I(n, 5)$ relative to $I(4, 3) = 100$, where (4, 3) is the strongest He II line in the optical region, at λ4686. Table 7 gives He I intensities relative to $I(3\,^3\mathrm{D},\, 2\,^3\mathrm{P}) = 100$.

Table 8 gives the effective recombination coefficients for Hβ, for He II λ4686 and for H I 3d–2p (from which the He I effective recombination coefficients are obtained using (2.32)). The values of $\alpha_{4,2}(\mathrm{H}^0)$ given in brackets are obtained allowing for l-degeneracy.

Table 5. Calculated Relative Intensities $I(n, 2)$ of Balmer Lines

	Case A		Case B	
T_e	1×10^4	2×10^4	1×10^4	2×10^4
n				
3	191 (248)	199 (236)	271 (262)	279 (253)
4	100 (100)	100 (100)	100 (100)	100 (100)
5	58·9 (50·1)	56·9 (51·0)	50·6 (48·9)	49·1 (49·3)
6	37·8	35·6	29·8	28·2
7	25·8	23·8	19·2	17·8
8	18·4	16·7	13·2	12·0
9	13·7	12·2	9·5	8·5
10	10·4	9·2	7·1	6·3

Table 6. He II Intensities (Case B) relative to $I(4, 3) = 100$ ($\lambda_{4,3} = 4686$)

	$I(n, 4)$		$I(n, 5)$	
T_e	1×10^4	2×10^4	1×10^4	2×10^4
n				
5	21·9	21·5	—	—
6	14·6	14·0	6·8	6·5
7	9·9	9·2	5·0	4·7
8	7·3	6·7	3·7	3·4
9	5·4	4·9	2·8	2·5
10	4·2	3·7	2·1	1·9

Table 7. He I Relative Intensities (calculated for $T_e = 1{\cdot}5 \times 10^4$ deg K)

n	3	4	5	6
$I(n\ ^3D, 2\ ^3P)$	100	41	21	12
$I(n\ ^1D, 2\ ^1P)$	29	12	6	4

Table 8. Effective Recombination Coefficients in $cm^3\ sec^{-1}$

	Case A		Case B	
T_e	1×10^4	2×10^4	1×10^4	2×10^4
$10^{14} \times \alpha_{4,2}$ (H^0)	1·35 (1·98)	0·73 (1·01)	3·00 (2·99)	1·60 (1·595)
$10^{14} \times \alpha_{4,3}$ (He^+)	9·0	5·2	20·8	11·8
$10^{14} \times \alpha_{3d,2p}$ (H^0)	5·22	2·35	5·27	2·38

2.6. *Comparisons with Observations*

The observed intensities must be corrected for reddening before a comparison with theory can be made (Shaïn 1934, Berman 1936). The available evidence indicates that the properties of the absorbing dust grains are the same in all parts of the galaxy and for each object we therefore have only one adjustable parameter, which determines the amount of interstellar dust in the line of sight. Let $I_0(\lambda)$ be the observed intensity of wavelength λ and $I(\lambda)$ the intensity which would be observed in the absence of dust. Then

$$\log I(\lambda) = \log I_0(\lambda) + cf(\lambda) \qquad \ldots\ldots(2.33)$$

where c is to be determined and $f(\lambda)$ is a known function. It is convenient to keep all relative intensities on the scale of $I(4861) = 100$ and we therefore take $f(\lambda)$ to be such that $f(4861) = 0$. If we also take $f(\lambda)$ to be such that $f(\infty) = -1$, the constant c in (2.33) is such that the absolute observed Hβ intensity must be multiplied by 10^c in order to correct for extinction. The function $f(\lambda)$ defined in this way has been obtained from the work of Whitford (1948, 1958) and of Divan (1954) and is given in table 9.

Intensities calculated according to the radiative recombination theory will be denoted by I_R. These are insensitive to T_e and to whether we assume Case A or Case B. We consider that Case B is the more likely to be correct for bright planetaries (§ 7.1) and that T_e will usually be in the range 1×10^4 to 2×10^4 deg K. We therefore consider only Case B and $T_e = 1{\cdot}5 \times 10^4$ deg K. We determine the reddening constant c by comparing I_0 and I_R for the stronger Balmer lines and then check that I and I_R agree for all of the H I, He I and He II lines. Table 10 gives results for the bright planetary NGC 7662, the observed intensities being taken from the work of Minkowski and Aller (1956). The agreement with theory is seen to be very satisfactory. This provides a check, not only on the validity of the theory, but also on the accuracy of the photometric measurements. Similar comparisons may be made for other planetaries but there are few for which the observational data are as complete. By far the most complete published list of wavelengths and intensities is that of Aller, Bowen and Minkowski (1955) for NGC 7027. For this nebula it is found that the reddening constant may be chosen so as to give agreement between I_R and I for all the stronger lines but there are systematic differences for the weaker lines (Seaton 1960). These are such that I_R/I_C tends to decrease with decreasing I_0 in the visual and ultra-violet regions and to increase with increasing λ in the infra-red. If it is believed that the radiative theory is correct—and on physical grounds it is difficult to see how it could be otherwise—then the differences between I_R and I must be due to systematic errors of measurement. One may then obtain empirical correcting factors, depending on I_0 and on λ, by which the measured intensities must be multiplied in order to obtain agreement between I_R and I.

2.7. *Determination of He^{+2} and He^+ Abundances*

The abundance ratios He^+/H^+ and He^{+2}/H^+ may be determined from

$$\left.\begin{aligned} \frac{N(He^+)}{N(H^+)} &= \frac{5876\,I(5876)}{4861\,I(4861)}\,\frac{\alpha_{4,2}(H^0)}{\alpha_{3^1D,\,2^1P}(He^0)}, \\ \frac{N(He^{+2})}{N(H^+)} &= \frac{4686\,I(4686)}{4861\,I(4861)}\,\frac{\alpha_{4,2}(H^0)}{\alpha_{4,2}(He^+)}. \end{aligned}\right\} \quad \ldots\ldots(2.34)$$

For Case B and $T_e = 1{\cdot}5 \times 10^4$ we have from table 7

$$\left.\begin{aligned} \frac{N(He^+)}{N(H^+)} &= 1{\cdot}00\,\frac{I(5876)}{I(4861)}, \\ \frac{N(He^{+2})}{N(H^+)} &= 0{\cdot}134\,\frac{I(4686)}{I(4861)}. \end{aligned}\right\} \quad \ldots\ldots(2.35)$$

Table 9. The Reddening Function $f(\lambda)$ ($f(4861) = 0$, $f(\infty) = -1$)

λ	$f(\lambda)$	λ	$f(\lambda)$
3000	0·65	6000	−0·26
3500	0·42	7000	−0·45
4000	0·24	8000	−0·60
4500	+0·10	10000	−0·81
5000	−0·04	20000	−0·97

Table 10. Intensities for NGC 7662

Series		n	I_0	I	I_R
H I	8863	11	6·2	3·2	1·53
$n \to 3$	8751	12	2·8	1·5	1·22
	8665	13	2·0	1·1	0·98
H I	6563	3	390	269	257
$n \to 2$	4861	4	*100*	*100*	*100*
	4340	5	44	51	49·8
	4102	6	23·2	28·7	29·0
	3835	9	7·7	10·3	9·0
	3798	10	5·7	7·8	6·7
	3771	11	4·6	6·3	5·0
	3750	12	3·8	5·2	4·0
	3734	13	3·4	4·7	3·2
	3712	15	2·2	3·1	2·1
	3697	17	1·5	2·1	1·4
	3692	18	1·3	1·8	1·2
	3687	19	0·9	1·3	1·0
	3683	20	0·73	1·0	0·9
He I	5876	3	16	*12·6*	*12·6*
$n\ ^3D \to 2\ ^3P$	4471	4	4·3	5·5	5·1
	4026	5	2·8	3·6	2·6
	3820	6	1·6	2·2	1·5
He I	6678	3	5·9	4·0	3·7
$n\ ^1D \to 2\ ^1P$	4922	4	1·3	1·3	1·5
	4388	5	0·6	0·7	0·8
	4144	6	0·33	0·4	0·4
He II	4686	4	51	*53*	*53*
$n \to 3$					
He II	5412	7	6·0	5·2	5·2
$n \to 4$	4542	9	2·4	2·6	2·8
	4200	11	1·4	1·7	1·6
	3858	17	0·62	0·8	0·5

Note: The intensities I_R have been adjusted so as to obtain agreement with I for the figures in italic type.

Table 11 gives some He abundances calculated from these expressions using intensity data of Aller (1941), Aller, Bowen and Minkowski (1955), Minkowski and Aller (1956), and Mathis (1957 c). We have selected objects for which one may check that the relative intensities in each spectrum (H I, He I and He II) are

in satisfactory agreement with theory and the abundance ratios should therefore be reliable. Adding the He^{+2} and He^{+} abundances we obtain an estimate of the total He/H abundance ratio but, if we assume Case B, we must consider the possibility that there is also some neutral He in the H^{+} region (see § 7).

Table 11. Helium Abundances

Nebula (NGC)	$\frac{N(He^{+2})}{N(H^{+})}$	$\frac{N(He^{+})}{N(H^{+})}$	$\frac{N(He^{+2})+N(He^{+})}{N(H^{+})}$
7622	0·071	0·126	0·197
2392	0·079	0·083	0·162
7027	0·067	0·127	0·194
6572	0·001	0·167	0·168
2022	0·151	0·050	0·201

Mathis (1957 c) has obtained He/H ratios for a larger number of planetaries and has shown that He/H shows no systematic variation with He^{+}/He^{+2}. Our He^{+} abundances are somewhat greater than those of Mathis (by $\sim 40\%$). Our He abundances are also greater than those of Nikitin (1950) (by a factor of ~ 2).

2.8. *Continuous Spectra of Planetaries*

Continuum observations have been made by Page (1936, 1942), Minkowski (1953), Barbier and Andrillat (1954) and Minkowski and Aller (1956). Electron–ion recombination and free–free transitions give important contributions to the continuum intensities but do not suffice to explain all of the observational data. Greenstein and Page (1951) have shown the H^{-} continuum to be unimportant. Kipper (1952) and Spitzer and Greenstein (1951) have considered the decay of the 2s metastable state of H with emission of two quanta: $H(2s) \rightarrow H(1s) + h\nu_1 + h\nu_2$ where $\nu_1 + \nu_2$ is the frequency of Lyα. Seaton (1955) has shown that the recombination, free–free and two-quantum continua give a satisfactory explanation of the observations provided that allowance is made for some collisional deactivation of H 2s.

For the continua the energy emitted in the interval $h\nu$ to $h\nu + dh\nu$ may be written

$$d(\text{energy}) = N_{+} N_{e} \gamma d h\nu$$

where γ is expressed in the units of a recombination coefficient ($cm^3\ sec^{-1}$). For recombination to level n of a hydrogenic ion

$$\gamma_n = \tfrac{1}{2} \mathcal{Q} Z \lambda^{1/2} \exp(-h\nu/kT) \left[\frac{2x_n \exp(x_n)}{n} g_{II}(n, \epsilon) \right] \quad \ldots\ldots (2.36)$$

in the notation of § 2.2. For free–free transitions,

$$\gamma_{ff} = \tfrac{1}{2} \mathcal{Q} Z \lambda^{1/2} \exp(-h\nu/kT) \bar{g}_{III} \quad \ldots\ldots (2.37)$$

where the mean Gaunt factor has asymptotic expansion

$$\bar{g}_{III} = 1 + 0{\cdot}1728 \left(\frac{h\nu}{Z^2 I_H} \right)^{1/3} \left\{ 1 + 2 \left(\frac{kT}{h\nu} \right) \right\}$$
$$- 0{\cdot}0496 \left(\frac{h\nu}{Z^2 I_H} \right)^{2/3} \left\{ 1 + \frac{2}{3} \left(\frac{kT}{h\nu} \right) + \frac{4}{3} \left(\frac{kT}{h\nu} \right)^2 \right\} + \ldots. \quad \ldots\ldots (2.38)$$

Let us define

$$\gamma = \sum_{n=n_0}^{\infty} \gamma_n + \gamma_{ff} \qquad \ldots\ldots(2.39)$$

where n_0 is the lowest level to which one can observe recombination at a given frequency. Tables 12 and 13 give $\log\gamma(H^0)$ and $\log\gamma(He^+)$, calculated using the expansions of the Gaunt factors. Within each interval the values of $\log\gamma$ may be interpolated as linear functions of frequency and should be correct to within 1%.

Table 12. Values of $\log_{10} 10^{14}\,\gamma(H^0)$

$h\nu/I_H$ \ $10^{-4}\,T_e$	0·8	1·0	1·2	1·4	1·6	1·8	2·0
0·30	1·295	1·247	1·200	1·156	1·116	1·081	1·049
0·25	1·709	1·575	1·471	1·387	1·317	1·258	1·207
0·25	0·176	0·320	0·408	0·465	0·504	0·531	0·552
0·16	0·931	0·922	0·909	0·894	0·879	0·865	0·853
0·1111	1·333	1·242	1·174	1·121	1·079	1·043	1·013
0·1111	0·799	0·815	0·821	0·821	0·818	0·813	0·808
0·10	0·890	0·890	0·883	0·875	0·865	0·856	0·847

Table 13. Values of $\log_{10} 10^{14}\,\gamma(He^+)$

$h\nu/I_H$ \ $10^{-4}\,T_e$	0·8	1·0	1·2	1·4	1·6	1·8	2·0
0·30	1·66	1·63	1·60	1·57	1·54	1·52	1·50
0·25	2·08	1·97	1·88	1·81	1·75	1·71	1·67
0·25	1·15	1·21	$1·24_5$	$1·26_5$	1·28	1·28	1·29
0·16	1·90	1·81	$1·74_5$	1·69	1·65	1·62	1·59
0·16	1·40	1·41	1·41	1·41	1·40	1·40	1·39
0·1111	1·80	1·73	1·68	1·64	1·61	1·58	1·56
0·1111	1·50	1·50	1·49	1·48	1·47	1·46	1·45
0·10	1·59	1·57	1·55	1·53	1·52	1·51	1·50

Table 14. Values of $10^{14}\,\gamma_0(2q)$ (Case B)

$h\nu/I_H$ \ $10^{-4}\,T_e$	0·8	1·0	1·2	1·4	1·6	1·8	2·0
0·30	10·59	8·97	7·83	6·99	6·33	5·81	5·41
0·25	8·54	7·23	6·31	5·63	5·10	4·68	4·36
0·20	6·43	5·45	4·76	4·25	3·85	3·53	3·29
0·15	4·35	3·69	3·21	2·87	2·60	2·39	2·22
0·10	2·43	2·06	1·80	1·61	1·45	1·33	1·24

Let $N(H^+)N_e\gamma(2q)\,d\boldsymbol{h\nu}$ be the energy in the two-quantum continuum. For densities so low that collisional deactivation of H 2s may be neglected we use the notation $\gamma_0(2q)$. Let $A(\boldsymbol{h\nu})$ be the probability per second that H 2s emits a photon in the range $d\boldsymbol{h\nu}$ and let A_{2q} be the total probability per second of $2s \rightarrow 1s$ transitions. Then

$$\int A(\boldsymbol{h\nu})\,d\boldsymbol{h\nu} = 2A_{2q} \qquad \ldots\ldots(2.40)$$

since each transition gives two photons; $A(\boldsymbol{h\nu})$ is tabulated by Kipper and by Spitzer and Greenstein and $A_{2q} = 8{\cdot}227\ \text{sec}^{-1}$. Defining $P(\boldsymbol{h\nu}) = A(\boldsymbol{h\nu})/2A_{2q}$ we have

$$\gamma_0(2q) = \boldsymbol{h\nu}P(\boldsymbol{h\nu})\,\alpha_{(2s)} \qquad \ldots\ldots(2.41)$$

where $\alpha_{(2s)} = X\alpha_B$ and where X, calculated neglecting collisions, has been given in table 4. Table 14 gives values of $10^{14}\gamma_0(2q)$ for Case B. The intensity of the

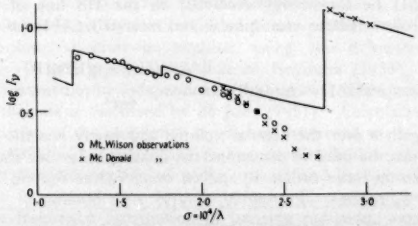


Figure 1. Continuum intensities in NGC 7027. The intensities I_ν per unit frequency, are in arbitrary units and $\sigma = 10^4/\lambda$ with λ in ångström units. The full line curve, calculated for $T_e = 1{\cdot}7 \times 10^4$ deg K, $N_e = 2{\cdot}2 \times 10^4\ \text{cm}^{-3}$ (see § 3.3) and the He abundances of table 11, is compared with the observations of Minkowski and Aller (1956).

two-quantum continuum is reduced at high densities due to collisional $2s \rightarrow 2p$ transitions. The rate of collisional deactivation, which is largely determined by proton impacts, has been calculated quite accurately. Seaton (1955) obtains finally

$$\gamma(2q) = \frac{\gamma_0(2q)}{1 + 0{\cdot}6 \times 10^{-4} N_e} \qquad \ldots\ldots(2.42)$$

for $T_e \simeq 10^4$ deg K. The maximum intensity of the H I two-quantum continuum occurs at $\nu = \frac{3}{4}\nu(\text{Ly}\alpha)$. Two-quantum continua also exist for He I and He II but they will be of little importance in the observable spectral region.

The total continuum intensity, per unit frequency, is proportional to

$$N(H^+)\{\gamma(H^0) + \gamma(2q)\} + N(He^+)\gamma(He^0) + N(He^{+2})\gamma(He^+) \qquad \ldots\ldots(2.43)$$

where, to a first approximation, one may put $\gamma(He^0) = \gamma(H^0)$. Page found the continuum intensities per unit wavelength to be approximately constant in the range $3647 \leqslant \lambda \leqslant 5000$ Å. This is in general agreement with theory. It has been shown by Seaton (1955) and by Aller (1956, p. 149) that there are no systematic differences between theory and observation for the ratios of continuum intensities to hydrogen line intensities.

The Balmer discontinuity Δ is defined as the logarithm of the ratio of the intensities on the two sides of the Balmer discontinuity at 3647 Å. Seaton (1955) has compared measured values of Δ with those calculated using the best available estimates of T_e, N_e. In § 4.3 we show how measured values of Δ may be used, in conjunction with other data, for the determination of T_e and N_e. Minkowski and Aller have measured the continuum intensities in NGC 7027 over an extended wavelength range. The observed intensities have been corrected for reddening and for an apparent calibration error in the infra-red (see § 2.6) and are compared with theory in figure 1. The agreement with theory is good throughout much of the range but is poor in the region to the immediate long wave side of the Balmer discontinuity, where the intensities are weak and the observational error may be large.

2.9. *The Absolute Brightness*

Let $\mathscr{I}$ (4861) be the energy received† in the Hβ line of the nebula in erg cm^{-2} sec^{-1}, corrected for atmospheric and interstellar absorption. Then

$$\mathscr{I}(4861) = h\nu(4861) \times \frac{\int N_e N(\mathrm{H}^+)\, \alpha_{4,2}(\mathrm{H}^0)\, dV}{4\pi r^2} \qquad \ldots\ldots(2.44)$$

where the integral is over the nebular volume and where r is the distance of the nebula. This may be used to determine the density. Assume the density to be uniform between an inner radius R_i and an outer radius R_o and let the angular radii be θ_i and θ_o radians. Putting $N_e = N(\mathrm{H}^+)$ we obtain

$$N_e^2 = 3\mathscr{I}(4861)/h\nu(4861)\,\alpha_{4,2}(\mathrm{H}^0)(\theta_o^3 - \theta_i^3)r. \qquad \ldots\ldots(2.45)$$

Accurate photoelectric measurements of $\mathscr{I}$ have been made by Liller and Aller (1954) and by Liller (1955). It is sometimes difficult to define the inner and outer radii but an improvement can be made by studying the variation of brightness across the disc (Wilson and Aller 1951). In using (2.45) the most uncertain quantity may be the distance r. Distances of planetaries are determined by a method depending on measured radial velocities and the theory of galactic rotation. While the values obtained should be statistically correct they may be very uncertain for individual objects. With accurate independent determination of N_e equation (2.45) could be used to estimate r. A further difficulty which arises is that, in addition to large-scale density variations, there may also be local condensation effects (§ 3.3).

2.10. *Collisional Excitation of Hydrogen*

We have considered the hydrogen levels to be populated by radiative recombination of H^+, the population rates being $N_e N_+ \alpha_n$. Another possibility, considered by Chamberlain (1953), is collisional excitation from the H ground state. The population rates would be $N_e N_1 \overline{v_1 Q_{1n}}$ where $N_1 = N(\mathrm{H}\,1s)$, where v_1 is the initial electron velocity, Q_{1n} the excitation cross section and $\overline{v_1 Q_{1n}}$ is averaged over a Maxwell distribution. With $T_e \simeq 10^4$ deg K, which is well established for

† Menzel and Aller (1941 b) use the notation S for the energy in erg cm^{-2} sec^{-1} at the surface of the nebula; evidently $\mathscr{I} = \theta^2 S$ where θ is the angular radius in radians.

planetaries, kT_e is small compared with the excitation energies $E_n - E_1$ and therefore $v_1 \overline{Q}_{1n}$ will be small and collisional excitation will be important only if N_1 is large. Thus, taking $T_e = 1{\cdot}4 \times 10^4$ deg K, Aller and Minkowski (1956) find that collisional excitation will be important in NGC 7027 only if $N_1 \sim \frac{1}{2}N_e$. Such a large abundance of neutral hydrogen is difficult to reconcile with the high degree of ionization shown by other elements and it is difficult to see how this condition could be maintained since $N_1 \sim \frac{1}{2}N_e$ would imply an optical depth of order 10^4 in the Lyman continuum, so that ionizing radiation could not penetrate.

We conclude that collisional excitation is not important in planetaries, although it may be important in certain other nebulae which have different energy sources and higher electron temperatures.

§ 3. Transfer Problems

3.1. *The Lyman Continuum*

Ambarzumian (1932 a, b) first considered the problem of the transfer of Lyman continuum radiation in nebulae, using the Schuster–Schwarzschild approximation. In this work, as in that of Hagihara (1938), the absorption coefficient was assumed to be independent of frequency. The question of effective absorption coefficients is discussed by de Jong (1951). Calculations allowing for the energy variation of the absorption coefficient have been made by Aller, Baker and Menzel (1939). Changes in the frequency distribution of the continuum radiation result from absorption of stellar quanta and emission of Lyman continuum quanta in ground-state recombinations. Electron temperatures, which depend on the frequency distribution of ionizing radiation, were calculated by Aller, Baker and Menzel for various depths in the nebula but these temperatures are too large due to neglect of the energy loss in forbidden line excitation (§ 9).

Wurm and Singer (1952) have considered the effect of He absorption on the continuum energy distribution.

3.2. *The H I Lyα Line*

For an optically thick nebula the flux of Lyα quanta is comparable with the flux of Lyman continuum quanta. If the Lyα quanta are scattered without change of frequency the absorption coefficient for Lyα exceeds that for the Lyman continuum by a factor of order 10^4 (§ 2.1) and the force exerted on the nebular material due to Lyα radiation pressure becomes very large, much greater than that due to gravitational attraction by the central star (Ambarzumian 1932 a, Chandrasekhar 1935). Zanstra (1934) has shown that this would lead to expansion velocities much greater than those observed. The first attempts to resolve this difficulty (Zanstra 1934, Chandrasekhar 1945) considered the reduction in Lyα radiation pressure due to Doppler effects which occur once expansion begins. A further effect, which may be more important, depends on the fact that each scattering process involves a change of frequency (Zanstra 1949, Sobolev 1944, 1955). Consider a photon with frequency ν propagated in the Oz direction and let ν_1 be the frequency of the line centre. This photon can be absorbed by an atom having velocity component $v_z = (c/\nu_1)(\nu - \nu_1)$ in the Oz direction. Subsequently the atom re-emits a photon and, since there is no restriction on the possible values

of the components of atom velocity, v_x and v_y, perpendicular to Oz, and since the photon may be emitted in any direction, the frequency of the re-emitted photon will differ from that of the absorbed photon. A first rough approximation is to assume that the probability of the photon being emitted with frequency ν' is proportional to $\exp[-b(\nu'-\nu_1)^2]$ (see § 2.1). The resulting frequency redistribution gives a greatly increased probability that the photon can escape in the wings of the line, where the absorption coefficient is small, and reduces the calculated radiation pressure by a factor of the order of 1/300 (Zanstra 1949). Further calculations taking account of redistribution have been made by Miyamoto (1950), Unno (1950, 1952 a, b), Zanstra (1951), Saigusa (1951) and Koelbloed (1956).

Most Lyα transfer calculations have been made for models having optical depths of order unity for the Lyman continuum. These models have little neutral hydrogen outside the ionized region (the boundary of the region in which most of the hydrogen is ionized occurs at an optical depth of order unity). Many nebulae may be surrounded by much more extensive neutral hydrogen regions. The possibility may be considered that Lyα absorptions could give large total H 2p populations in such regions and hence large optical depths for Balmer quanta. Observed spectra suggest that planetaries are optically thin for Balmer quanta (§ 2.2) and it is therefore of interest to consider the processes which result in Lyα quanta being destroyed.

One possibility is that $2p \rightarrow 2s$ transitions produced by collisions are followed by two-quantum emission. Since the probability of $2p \rightarrow 2s$ transitions is small at nebular densities, compared with the probability of Lyα emission, such processes are important only if each Lyα quantum is scattered very many times. It has been shown by Spitzer and Greenstein (1951) and by Yada (1953, 1954) that such processes are not important in H^+ regions but could be important in sufficiently extensive H^0 regions. The efficiency of this process in H^0 regions is reduced by the fact that $2p \rightarrow 2s$ transitions are produced most readily by charged particle impacts and that the only charged particles in H^0 regions are those resulting from ionization of elements such as C and Fe of low ionization potentials.

Seaton (1955) has suggested that a faint outer envelope observed by Duncan (1937) for NGC 6826 might be due to the two-quantum continuum produced in an H^0 region but R. Minkowski (private communication) has since found that the envelope has a typical nebular line spectrum.

The problem of the transfer of Lyα radiation in extensive H^0 regions has been treated in detail by Yadi and Ōsaki (1957) who show that absorption by dust grains will be very effective in destroying much of the Lyα radiation if the dust concentration, relative to that of hydrogen, has a value comparable with that for typical interstellar clouds. It is, of course, possible that the H^0 envelope surrounding a planetary nebula might have a much smaller dust concentration.

3.3. *The He II Lyα Line*

Many nebulae will have a He^{+2} region surrounded by a He^+ region. Unno (1952 c, 1955) has treated the problem of He II Lyα transfer using methods similar to those which have been employed for H I Lyα. The two cases differ in that He II Lyα is destroyed more readily than H I Lyα. Due to the higher charged particle density, conversion to two-quanta emission will be more effective for

He II Lyα in a He^+ region than for H I Lyα in an H^0 region. For He II Lyα a more important loss process will probably be photo-ionization of H^0 and He^0. In some nebulae an appreciable fraction of all H^0 ionizations will be produced by He II Lyα quanta (see § 8).

§ 4. The Forbidden Lines

Most of the observed forbidden lines are due to transitions within the ground configurations of atoms or ions with outer $2p^q$ or $3p^q$ electrons, where $q = 2$, 3 or 4. These configurations have three spectral terms denoted by $n = 1$, 2 and 3 in order of increasing excitation energy. They are

n	$q = 2$ and 4	$q = 3$
3	^1S	^2P
2	^1D	^2D
1	^3P	^4S

The transition probabilities for the observed lines, calculated for electric quadrupole and magnetic dipole interactions, range from 10^{-5} to 10 sec^{-1}. A list is given by Aller (1956) and a review by Garstang (1957).

4.1. *Excitation Mechanisms*

Bowen (1928) suggested that the forbidden lines were excited by electron impact and this had been confirmed by all subsequent work; other processes, such as proton impact and recombination, have been shown to be unimportant (Seaton 1954 a, Burgess and Seaton 1960 b).

The cross sections for electron impact are conveniently written

$$Q(n \to n') = \frac{\pi \Omega(n, n')}{k_n^2 \omega_n} \qquad \ldots\ldots(4.1)$$

(Hebb and Menzel 1940) where $k_n = mv_n/\hbar$ (v_n being the initial electron velocity) and ω_n is the statistical weight of level n. The *collision strength* $\Omega(n, n')$ is dimensionless and is symmetrical in n, n'. The probability per sec of collisional *deactivation* is

$$d_{n'n} = \frac{8{\cdot}63 \times 10^{-6} N_e}{\omega_n T_e^{1/2}} \overline{\Omega}(n, n') \quad (n' > n) \qquad \ldots\ldots(4.2)$$

where

$$\overline{\Omega}(n, n') = \int_0^\infty \Omega(n, n') \exp\left(-\frac{mv_{n'}^2}{2kT_e}\right) d\left(\frac{mv_{n'}^2}{2kT_e}\right). \qquad \ldots\ldots(4.3)$$

For positive ions Ω is practically independent of energy and one may put $\overline{\Omega} = \Omega$. For *excitation*

$$d_{nn'} = \frac{\omega_{n'}}{\omega_n} d_{n'n} \exp\left(-\frac{E_{n'} - E_n}{kT_e}\right) \quad (n' > n). \qquad \ldots\ldots(4.4)$$

Values of collision strengths are given in a review by Seaton (1958). For planetaries it is convenient to introduce the parameters x and t defined by

$$x = \frac{10^{-4} N_e}{t^{1/2}}, \quad t = 10^{-4} T_e. \qquad \ldots\ldots(4.5)$$

These will be of order unity.

4.2. *Equilibrium Equations and Intensity Ratios*

Consider first an excited state for which the radiative transition probability is much greater than the collisional deactivation probability. The population will be proportional to $N_e N_1 d_{1n}$. If an ion has two such states the relative intensities will be independent of N_e and will depend on T_e only through an exponential factor, $\exp(-\Delta E/kT_e)$. Such ratios may be used to determine T_e. If, on the other hand, the collisional deactivation probability is large compared with the radiative probability the ratio N_n/N_1 will approach the value for thermodynamic equilibrium. If this applies to two excited states the relative intensities will again be independent of N_e. Now consider that an ion has two excited states, such that the radiative probabilities are large for one and small for the other. The intensity ratios may then be sensitive to N_e and provide a useful method of determining densities.

We consider the quantitative theory. Total transition probabilities are

$$\left.\begin{aligned} p_{12} &= d_{12}, & p_{13} &= d_{13}, \\ p_{21} &= d_{21}+A_{21}, & p_{23} &= d_{23}, \\ p_{31} &= d_{31}+A_{31}, & p_{32} &= d_{32}+A_{32} \end{aligned}\right\} \qquad \ldots\ldots(4.6)$$

where the $A_{nn'}$ are radiative probabilities. The steady state equations are

$$\sum_{n'} N_{n'} p_{n'n} = N_n p_n \qquad \ldots\ldots(4.7)$$

where $p_n = \Sigma_{n'} p_{nn'}$. The solutions may be written

$$N_n = \frac{P(n)\,N}{S} \qquad \ldots\ldots(4.8)$$

where $S = \Sigma_n P(n)$ and $N = \sum_n N_n$. For three levels,

$$\left.\begin{aligned} bP(1) &= p_2 p_3 - p_{32} p_{23} \\ bP(2) &= p_{12} p_3 + p_{13} p_{32} \\ bP(3) &= p_{13} p_2 + p_{12} p_{23} \end{aligned}\right\} \qquad \ldots\ldots(4.9)$$

where b is any constant.

The solutions for O^{+2} are

$$\left.\begin{aligned} P(^3\mathrm{P}) &= 1+0{\cdot}010x(1+0{\cdot}05\epsilon)+3{\cdot}9\times10^{-6}x^2(1+0{\cdot}10\epsilon) \\ P(^1\mathrm{D}) &= 5{\cdot}4\times10^{-3} \\ &\quad\times\exp(-2{\cdot}89/t)\,x\{1+0{\cdot}12\epsilon+4{\cdot}1\times10^{-4}x(1+0{\cdot}11\epsilon)\} \\ P(^1\mathrm{S}) &= 1{\cdot}16\times10^{-5}\exp(-6{\cdot}19/t)\,x\{1+0{\cdot}038x(1+0{\cdot}10\epsilon)\} \end{aligned}\right\} \qquad \ldots\ldots(4.10)$$

where $\epsilon(O^{+2}) = \exp(-3{\cdot}30/t)$. For x and t of order unity we obtain the [O III] intensity ratio

$$\frac{I(4959)+I(5007)}{I(4363)} = \frac{7{\cdot}1\times\exp(3{\cdot}30/t)}{1+0{\cdot}038x}. \qquad \ldots\ldots(4.11)$$

For $x\sim1$ a good first approximation is to neglect the term $0{\cdot}038x$ and to use the ratio (4.11) to determine T_e.

For [N II] one obtains

$$\frac{I(6548)+I(6584)}{I(5755)} = \frac{6{\cdot}3\times\exp(2{\cdot}50/t)}{1+0{\cdot}24x}, \qquad \ldots\ldots(4.12)$$

which is more sensitive to x.

We have assumed the fine structure levels, ^3P_J, of the O^{+2} and N^+ ground states to be populated in the statistical weight ratios $2J+1$. Using collision strengths and transition probabilities for transitions between the O^{+2} fine structure levels (Hebb and Menzel 1940, Seaton 1953) it may be shown that this is a fair approximation but is not exact. However, so far as the excited term populations are concerned, no error is introduced; this is because $\Omega(^3P_J, {}^1D)$ and $\Omega(^3P_J, {}^1S)$ are both proportional to $2J+1$.

For ions such as O^+, with $q=2$, fine structure of the ^2D and ^2P levels should be taken into account in order to obtain accurate expressions for the relative intensities. The functions P for O^+ ^4S, $^2D_{5/2}$, $^2D_{3/2}$, $^2P_{3/2}$ and $^2P_{1/2}$ are given by Seaton and Osterbrock (1957). For the intensity ratio $(^2D\rightarrow{}^4S)/(^2P\rightarrow{}^2D)$ one obtains

$$\frac{I(3729)+I(3726)}{I(7320)+I(7330)} = \frac{5{\cdot}5}{\epsilon}\left[\frac{1+0{\cdot}36\epsilon+5{\cdot}3x(1+0{\cdot}82\epsilon)}{1+13{\cdot}8x(1+0{\cdot}38\epsilon)+38{\cdot}4x^2(1+0{\cdot}78\epsilon)}\right] \qquad \ldots\ldots(4.13)$$

where $\epsilon(O^+) = \exp(-1{\cdot}96/t)$ and for $(^2D_{5/2}\rightarrow{}^4S)/(^2D_{3/2}\rightarrow{}^4S)$ one obtains

$$\frac{I(3729)}{I(3726)} = 1{\cdot}5\left[\frac{1+0{\cdot}33\epsilon+2{\cdot}30x(1+0{\cdot}75\epsilon)}{1+0{\cdot}40\epsilon+9{\cdot}9x(1+0{\cdot}84\epsilon)}\right]. \qquad \ldots\ldots(4.14)$$

In most planetaries both of the [O II] ratios, (4.13) and (4.14), are sensitive to N_e and provide methods of determining densities (the relative intensities within the $^2P\rightarrow{}^2D$ multiplet, on the other hand, are insensitive to T_e and N_e). The intensity ratios for [S II] give relations similar to those for [O II] but the atomic data for S^+ are a good deal less accurate than those for O^+.

4.3. *Determination of Temperatures and Densities*

In general each observed intensity ratio gives a relation between T_e and N_e. Seaton (1954 b) has obtained temperatures and densities on making plots of T_e against N_e for a number of planetaries. Figure 2 shows a number of such plots, recalculated using more recent atomic and observational data.† In addition to the relations obtained from the forbidden lines we include those from the absolute surface brightness (using data quoted by Aller 1956, p. 149) and from the Balmer discontinuity (using measurements by Barbier and Andrillat (1954) and the theory of § 2·8). Aller (1954, 1957) has obtained T_e and N_e from simultaneous solutions of the [N II] and [O III] equations. It is seen from figure 2 that the [N II] and [O III] curves cross at an angle which is too small for this method to give a good density estimate.

So far as close orders of magnitude are concerned the various methods are seen to give good agreement for each nebula. Closer inspection shows that there are some quite systematic differences. The density obtained from the surface

† I am indebted to Dr. D. E. Osterbrock for an unpublished 3729/3726 intensity ratio for NGC 2440.

brightness is in all cases greater than that obtained from the 3729/3726 [O II] ratio and is usually less than that obtained from other forbidden line ratios and from the Balmer discontinuity. These differences are too large to be due to errors in the observations or in the atomic data. They can be explained if one gives up the idea that there is a unique temperature and density for each object (Seaton and

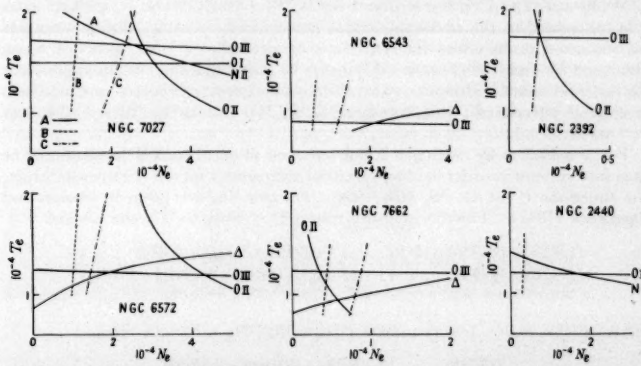


Figure 2. (T_e, N_e) relations for six planetary nebulae. A : from the Balmer discontinuity Δ and the relative intensities of forbidden lines ; B : from the [O II] 3729/3726 intensity ratio ; C : from the absolute surface brightness in Hβ.

Osterbrock 1957). For the moment let us only consider possible local density variations. The simplest model is that which supposes dense clouds to be embedded in a medium of lower density. The 3729, 3726 lines have very small transition probabilities and compared with other lines are strongest in the low density regions. Therefore the effective density determined from the 3729/3726 ratio will be low. The surface brightness method gives a root mean square density

$$N_e(\text{rms}) = \left\{ V^{-1} \int N_e^2 \, dV \right\}^{1/2}. \qquad \ldots\ldots(4.15)$$

It may be shown that the other forbidden line ratios will usually give densities greater than N_e(rms). Seaton and Osterbrock consider the total volume V to be divided into two volumes V_1 and V_2 having densities N_{e1} and N_{e2}. For NGC 7027 they estimated $N_{e1} \simeq 5 \times 10^3\,\text{cm}^{-3}$, $N_{e2} \simeq 5 \times 10^4\,\text{cm}^{-3}$ and $V_2/V_1 = 0{\cdot}01$ to 0·02.

A theory which may explain the formation of local condensations will be considered in §9.

§5. Fluorescent Mechanisms

Quantitative discussions of fluorescent mechanisms have been given by Menzel and Aller (1941 a,), by Hatanaka (1944), by Unno (1955) and by Burgess and Seaton (1960 b).

5.1. *The O III Lines*

The He II Lyα line has $\lambda = 303{\cdot}780$ and the O III line $2p^2\,{}^3P_2$–$2p\,3d\,{}^3P_2$ has $\lambda = 303{\cdot}799$. Since He II Lyα will be broad the coincidence in wavelength may be considered almost exact. Following excitation to 2p 3d an O^{+2} ion may return directly to the ground state or it may go to 2p 3p and then return to the ground state via 2p 3s. The 3d$\rightarrow$3p and 3p$\rightarrow$3s lines, many of which are in the observable ultra-violet, are known as fluorescent lines.

The intrinsic efficiency of the fluorescent mechanism is defined as the probability of fluorescent lines being emitted following a single He II Lyα absorption. This efficiency is low, being only 0·017 according to Burgess and Seaton. The overall efficiency is defined as

$$\mathscr{R}(\text{O III}) = \frac{\mathscr{N}(\text{O III}, 3d \rightarrow 3p)}{\mathscr{N}(\text{He II}, \text{Ly}\alpha)}$$

where $\mathscr{N}(\text{O III}, 3d \rightarrow 3p)$ is the number of 3d$\rightarrow$3p quanta emitted by the nebula and $\mathscr{N}(\text{He II}, \text{Ly}\alpha)$ is the number of He II Lyα quanta produced in the nebula. Despite its low intrinsic efficiency the fluorescent mechanism may have a high overall efficiency if each He II Lyα quantum is absorbed many times. The overall efficiency may be determined from the relative intensities in the nebular spectrum.

We consider the fluorescent line $3d\,{}^3P_2 \rightarrow 3p\,{}^3P_2$, $\lambda 3444$, since it is strong and is not too close to the limit of atmospheric transmission. From the cascade theory of Burgess and Seaton $\mathscr{N}(\text{O III}, \lambda 3444) = 0{\cdot}264\mathscr{N}(\text{O III}, 3d \rightarrow 3p)$. The number of He II Ly$\alpha$ quanta is

$$\mathscr{N}(\text{He II}, \text{Ly}\alpha) = (1 - X)\,\alpha_B(\text{He}^+)\int N_e\,N(\text{He}^{+2})\,dV$$

(§ 2.8). For some observed He II line

$$\mathscr{N}(\text{He II}, n \rightarrow n') = \alpha_{nn'}(\text{He}^+)\int N_e\,N(\text{He}^{+2})\,dV.$$

We choose the (5, 3) He II line at $\lambda 3202$ since it is close to the O III lines (it may be checked that the observed intensities corrected for reddening are consistent with the theory relation $I_R(3202) = 0{\cdot}56\ I_R(4686)$). We obtain finally

$$\mathscr{R}(\text{O III}) = 0{\cdot}30\,\frac{I(\text{O III}, \lambda 3444)}{I(\text{He II}, \lambda 3202)},$$

where the numerical factor is calculated with $X = 0{\cdot}3$ which will be approximately correct if collisional He^+ 2s–2p transitions can be neglected. Table 15 gives values of $\mathscr{R}$(O III) calculated using the intensities of Aller (1951).

Table 15. Overall Efficiencies of the O III and N III Fluorescent Mechanisms

Nebula	$\mathscr{R}$(O III)	$\mathscr{R}$(N III)
IC 351	0·26	0·34
IC 2003	0·31	0·33
NGC 2165	0·33	0·43
J 900	0·33	0·29
NGC 2440	0·22	1·1
NGC 3242	0·49	0·55
An. $21^h\ 31^m$	0·10	0·34

5.2. *The N III Lines*

The line O III $2p\,3s\,^3P_1 \rightarrow 2p^2\,^3P_2$, λ373·436, produced by the fluorescent mechanism, has a wavelength coinciding almost exactly with those of the N III lines $2p\,^2P_{3/2}$–$3d\,^2D_{5/2}$, λ374·434 and $2p\,^2P_{3/2}$–$3d\,^2D_{3/2}$, λ374·442. From N III 3d transitions can take place to 3p (giving N III, λ4634, 4640 and 4642) and then to 3s and to the ground state. The intrinsic efficiency for the N III fluorescent mechanism is even lower than that for O III, being only 0·005 (this is calculated by a method similar to that used by Burgess and Seaton for O III). The overall efficiency is

$$\mathscr{R}(\text{N III}) = \frac{\mathscr{N}(\text{N III},\ 3d \rightarrow 3p)}{\mathscr{N}(\text{O III},\ \lambda 374{\cdot}436)}$$

where $\mathscr{N}$(O III, λ374·436) is the number of λ374·436 quanta produced in the nebula. From the O III cascade theory, $\mathscr{N}$(O III, λ374·436) = 0·129$\mathscr{N}$(O III, 3d→3p). We obtain for the overall efficiency

$$\mathscr{R}(\text{N III}) = 2{\cdot}75\,\frac{I(\text{N III},\ \lambda 4634, 4641, 4642)}{I(\text{O III},\ \lambda 3444)}.$$

Numerical values are given in table 15. Our values of $\mathscr{R}$(N III) will be overestimated if recombination contributes to the N III line intensities.

5.3. *Discussion of the Fluorescent Efficiencies*

The efficiencies $\mathscr{R}$(O III) and $\mathscr{R}$(N III) are seen to be about 0·3 in most nebulae. For a single scattering process of optical depth τ the number of scatterings is of order τ^2. Burgess and Seaton show that the optical depths will be of order 10^4 for the centre of the O III absorption line assuming thermal Doppler broadening. Similar optical depths would be obtained for the N III absorption line, and the central optical depths for He II Lyα will be of order 10^6–10^7. It would be an interesting problem to obtain a quantitative theory of the fluorescent efficiencies, which are not very close to unity despite the fact that the number of scatterings would appear to be very large. In addition to the effect of the broadening of the line contours by incoherence (§ 3) one should also consider other processes which reduce the number of quanta available for excitation of fluorescent lines. The He II Lyα and the O III resonance lines may be absorbed in ionization of H^0 and He^0. An additional loss process for He II Lyα has already been mentioned in §3.3; collisional 2p→2s transitions in He^+ may be followed by 2s→1s two-photon emission.

§ 6. Chemical Compositions

Since nebulae are optically thin for the observed lines the ionic abundances are directly proportional to the line intensities. In this respect the problem of determining the composition of a nebula is simpler than that of determining the composition of a stellar atmosphere but the nebular problem is the more difficult one in that conditions are very far from those of thermodynamic equilibrium. The relative abundances of various ions may be obtained when all relevant cross sections and transition probabilities are known. The largest source of error in the determination of chemical compositions arises from the fact that most elements are present in many different stages of ionization and that the abundances can usually be determined observationally for only one or two stages.

The first estimates of chemical compositions were made by Bowen and Wyse (1939) and Wyse (1942) while the first attempt at a detailed quantitative theory was made by Aller and Menzel (1945). Subsequent work has been done by Aller (1953 b), Seaton (1954 b), Aller (1954), Wurm (1954), Minkowski and Aller (1956), Aller (1957), Mathis (1957 c), Seaton (1960) and Burgess and Seaton (1960 b).

6.1. *Ionic Abundances from Recombination Lines*

We consider ionic abundances relative to H^+ abundances since these are more accurate than absolute ionic abundances. Since the observed lines are produced in the ionized part of a nebula the H^+ concentration is equal to the total hydrogen concentration. Let $I(\lambda)$ be a line intensity relative to 100 for Hβ, λ4861, and define the relative quantum emission rate as

$$q(\lambda) = \left(\frac{\lambda}{4861}\right) I(\lambda). \qquad \ldots\ldots(6.1)$$

For a recombination line λ_{nm} in the spectrum of an ion X^{+p} the emission rate $q(X^{+p}, \lambda_{nm})$ is proportional to $N_e N(X^{+p+1}) \alpha_{nm}(X^{+p})$ where α_{nm} is the effective recombination coefficient (§2.2). Since $q(H\beta) = 100$,

$$\frac{q(X^{+p}, \lambda_{nm})}{100} = \frac{N(X^{+p+1})\,\alpha_{nm}(X^{+p})}{N(H^+)\,\alpha_{4\beta}(H^0)}. \qquad \ldots\ldots(6.2)$$

Relations of this type have been used in §2.7 for the determination of He/H abundance ratios. A method of calculating recombination spectra for complex ions has been given by Burgess and Seaton (1960 b). The method of Bates and Damgaard (1949) is used for the calculation of radiative transition probabilities and a further development of this method (Burgess and Seaton 1960 a) is used for the calculation of photo-ionization cross sections and recombination coefficients. From calculated recombination spectra for O III, O IV and O V abundances of O^{+3}, O^{+4} and O^{+5} have been obtained.

As in the case of hydrogenic ions (§2) we have to consider whether Case A or Case B should be used. For many excited states of complex ions the probability of a transition to the ground state is much greater than the probability of a transition to a lower excited state and in consequence some of the recombination lines calculated for Case B are much stronger than those calculated for Case A. In such circumstances the quanta in the resonance lines must be scattered many times in order that Case B should hold. The conditions for which Case B will be a good approximation are similar to those required to give overall efficiencies close to unity for the fluorescent mechanisms. Although the optical depths will usually be large for the resonance lines it follows from the discussion of §5.3 that this does not necessarily ensure that Case B is a good approximation.

For NGC 7027, O^{+3} is found to be the most abundant oxygen ion. Fortunately the calculated O^{+3} abundance is not sensitive to which Case is assumed. The abundances calculated for O^{+4} and O^{+5} are much more sensitive; the assumption of Case B may lead to $N(O^{+4})$ and $N(O^{+5})$ being underestimated by a factor of about two or three.

6.2. *Ionic Abundances from Forbidden Lines*

For a forbidden line the emission rate $q(X^{+p}, \lambda_{nm})$ is proportional to $A_{nm} N_n(X^{+p})$ where $N_n = P(n) N/S$ (§4.2). The abundance ratios are calculated from

$$\frac{q(X^{+p}, \lambda_{nm})}{100} = \frac{N(X^{+p}) A_{nm} P(X^{+p}, n)}{N(H^+) S(X^{+p}) N_e \alpha_{4\beta}(H^0)}. \qquad \ldots\ldots(6.3)$$

The abundance ratios are usually sensitive to T_e but if the radiative transition probabilities are large compared with collisional deactivation probabilities, $P(n)$ will be nearly proportional to N_e and S will be nearly independent of N_e, so that $N(X^{+p})/N(H^+)$ will be insensitive to N_e.

6.3. *Determination of Total Abundances*

The ionization equilibrium is determined by

$$N(X^{+p}) \int a_\nu(X^{+p}) \frac{c}{h\nu} \rho_\nu(\text{neb})\, d\nu = N_e N(X^{+p+1}) \alpha(X^{+p}). \qquad \ldots\ldots(6.4)$$

The difficulty in applying this equation is that we have only very imperfect knowledge of ρ_ν(neb), the radiation density in the nebula. The ionization equilibrium for complex ions has therefore usually been treated in an approximate empirical manner. The procedure introduced by Aller and Menzel (1945) is to plot $(\log N(X^{+p}) + \text{constant})$ against $I(X^{+p-1})$ where $I(X^{+p})$ is the ionization potential of X^{+p}. The basic assumption is that, with suitable choice of the constant for each chemical element, all elements will give points lying on the same smooth curve. In practice we usually have no more than two or three observed ions for each element. Plots of $(\log N(X^{+p}) + \text{constant})$ against $I(X^{+p-1})$ using, for example, the data of Aller (1957), show a good deal of scatter.

6.4. *Results for* NGC 7027

To illustrate the methods of abundance determinations some results are obtained for NGC 7027. The intensities of Aller, Bowen and Minkowski (1955) are used with corrections mentioned in § 2.6 and all calculations are made with

Table 16. Ionic Abundances in NGC 7027

X^{+m}	$10^3 \frac{N(X^{+m})}{N(H^+)}$	X^{+m}	$10^3 \frac{N(X^{+m})}{N(H^+)}$
O^0	0·35	Ne^{+2}	3·2
O^+	1·0	Ne^{+3}	1·5
O^{+2}	15	Ne^{+4}	5·9
O^{+3}	64		
O^{+4}	28	N^0	0·10
O^{+5}	8	N^+	1·0

$T_e = 1{\cdot}7 \times 10^4$ deg K and $N_e = 2{\cdot}2 \times 10^4$ cm^{-3} (see figure 2). Table 16 gives the abundances of some ions with outer 2p electrons, for which the atomic data should be tolerably accurate. The abundances of O^{+3}, O^{+4} and O^{+5} are obtained from recombination lines, the values quoted for O^{+4} and O^{+5} being twice those calculated

assuming Case B (§ 6.1). All other abundances of table 16 are obtained from forbidden lines.†

Figure 3 gives ionization functions, $\log N(X^{+m}) + C(X)$ plotted against $I(X^{+m})$, the constants $C(N)$ and $C(Ne)$ being chosen so as to obtain approximate agreement with the smooth curve drawn through the oxygen points. Due to differences in the atomic constants (I, a_ν and α) in (6.4) we cannot expect to obtain exact agreement between the ionization functions for different elements but this is unlikely to be the main cause of the scatter of the N and Ne points in figure 3. In addition to errors in the atomic data and in the measured intensities, further important errors may arise from assuming T_e and N_e to be constant throughout the nebula.

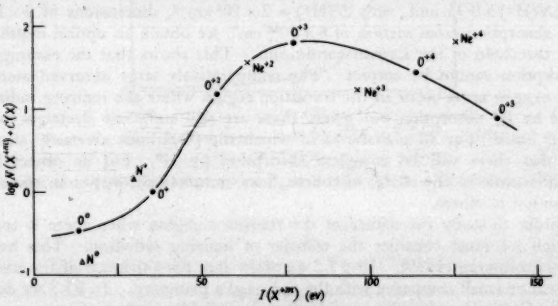


Figure 3. Ionization functions for NGC 7027.

Chemical abundances are given in table 17. For He, O and Ne the abundances are determined observationally for all stages of ionization giving significant contributions but the total N abundance is estimated using the ionization function

Table 17. Chemical Compositions

X	log N (X)			
	NGC 7027	10 Lac O_9 V	τ Sco B_0	ζ Per B_1 I
H	6·00	6·00	6·00	6·00
He	5·29	5·23	5·23	5·31
N	2·92	2·37	2·57	2·31
O	3·06	2·77	3·12	3·03
Ne	2·03	2·72	3·72	2·61

for oxygen. In table 17 the results for NGC 7027 are compared with those for the atmospheres of the early type stars 10 Lac (Traving 1957), τ Sco (Traving 1955) and ζ Per (Cayrel 1957). The composition of the nebula is seen to be similar to that of the stellar atmospheres, the agreement being particularly close for the ratios which are the most accurately determined, He/H and O/H.

† I am indebted to Dr. R. H. Garstang for supplying improved transition probabilities used in the Ne IV calculations.

§ 7. Sizes of Ionized Regions

7.1. *Optical Depths for H Absorption*

For many years there has been uncertainty as to whether there is complete absorption of Lyman continuum radiation by planetary nebulae. The difficulty is that it has not been possible to deduce the abundance of neutral hydrogen directly from observations. Fortunately it is possible to deduce the abundance of neutral oxygen from the [O I] lines and in many cases this enables us to settle the matter. From the data of § 6.4 we have $N(O^0)/N(O^+) = 0{\cdot}35$ for NGC 7027. The photo-ionization and recombination coefficients for O^0 and H^0 are of comparable magnitude. Let us assume that there is no absorption.† We then have $N(H^0)/N(H^+) \simeq 0{\cdot}35$ and, with $N(H^+) = 2\times10^4$ cm^{-3}, dimensions of 2×10^{17} cm and an absorption cross section of 6×10^{18} cm^2, we obtain an optical depth of 10^4 for the threshold of the Lyman continuum. This shows that the assumption of no absorption cannot be correct. The comparatively large observed amount of neutral oxygen must occur in the transition region where the ionizing radiation is reduced by H^0 absorption but where there are still sufficient electrons to excite the [O I] lines. For all planetaries in which the [O I] lines are fairly strong we expect that there will be complete absorption by H^0. For an object lacking spherical symmetry one could, of course, have complete absorption in some directions but not in others.

In order to study the nature of the transition region when there is complete absorption we must consider the transfer of ionizing radiation. This has been done by Strömgren (1939). In § 7.2 we show that the thickness of the transition region will be small compared with the radius of a planetary. In § 7.3 we calculate the radii of ionized regions by considering the available number of ionizing quanta, an approach similar to that employed by Zanstra (1931). The question of sizes of ionized regions has been discussed by Page and Greenstein (1951), Wurm and Singer (1952), Gurzadian (1952) and Van de Hulst (1952).

7.2. *The Transition Region*

We put $N_1 = N(H^0, 1s)$, $N_+ = N(H^+)$ and assume the total H density, $N = N_1 + N_+$, to be constant. Let $\mathcal{F}$ be the photon flux (cm^{-2} sec^{-1}) assumed monochromatic, a the photo-ionization cross section and α the recombination coefficient. With $N_e = N_+$ the ionization equation is

$$\mathcal{F}aN_1 = \alpha N_+^2. \qquad \ldots\ldots(7.1)$$

Neglecting ionizing quanta produced on recombination the transfer equation is

$$\frac{d\mathcal{F}}{dx} = -a\mathcal{F}N_1 \qquad \ldots\ldots(7.2)$$

for a plane parallel model. Substituting $N_1 = \xi N$, $N_+ = (1-\xi)N$ and $X = aNx$ we obtain $d\xi/dX = \xi^2(1-\xi)/(1+\xi)$ with solution

$$X(\xi) - X(\tfrac{1}{2}) = 2 - \frac{1}{\xi} + 2\ln\left(\frac{\xi}{1-\xi}\right). \qquad \ldots\ldots(7.3)$$

† Strictly speaking, our argument applies to a thin shell, since we do not allow for the variation of abundance ratios with position. But a more elaborate treatment would not alter our conclusions.

This is plotted in figure 4. The transition region is seen to have a thickness of $\Delta X \sim 20$ or, with $a = 10^{-17}$ cm^2 and $N = 10^4$ cm^{-3}, $\Delta x \sim 2 \times 10^{14}$ cm. This is about 10^{-3} of the radius of a typical planetary. In a spherically symmetric nebula variation of the geometrical dilution factor causes the transition region to be even sharper.

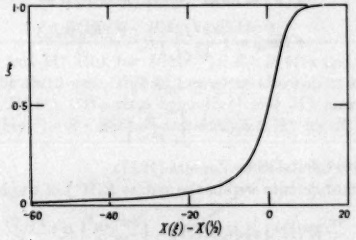


Figure 4. The transition region ($\xi = N(\mathrm{H}^0)/N(\mathrm{H})$ and $X = xaN(\mathrm{H})$).

Consider that $\xi = \xi_0$ at some point P_0 and that $\xi_0 \ll 1$. Then from (7.3),

$$X(\xi_0) - X(\tfrac{1}{2}) \simeq -\frac{1}{\xi_0} \quad (\xi_0 \ll 1). \qquad \ldots\ldots(7.4)$$

Let $X(\xi_0) - X(\frac{1}{2})$ correspond to a distance r_0. From (7.4), $aNr_0\xi_0 = 1$ and putting $N_+ = N$ in (7.1),

$$\mathscr{F}_0 = r_0 \alpha N^2. \qquad \ldots\ldots(7.5)$$

This equation has the following interpretation. We have a completely ionized region with a sharp boundary at a distance r_0 from the point P_0. Since $N_+ = N$ in the ionized region, $r_0 \alpha N^2$ is the number of recombinations between P_0 and the boundary and this is equal to the photon flux $\mathscr{F}_0$ at P_0. In so far as a sharp boundary may be assumed, equation (7.5) is exact.

7.3. *Radii of Ionized Regions*

Assuming that all ionizing quanta are absorbed and that a sharp boundary exists, we determine the radii of ionized regions on equating the number of ionizations to the number of recombinations. Ionizations may be due to stellar quanta or to quanta produced in the nebula. Since planetaries are thin in Balmer and higher continua (§2.2), ionizing quanta will be produced only by ground state recombinations. We may therefore equate the number of ionizing quanta produced in the nebula to the number of ground state recombinations and the number of ionizing stellar quanta to the number of excited state recombinations.

Let $\mathscr{N}(h\nu_0)$ be the number of quanta, with energy greater than $h\nu_0$, emitted by the star. The size of the H^+ region is determined from the equation

$$\mathscr{N}(I_{\mathrm{H}}) = \int N_e N(\mathrm{H}^+) \alpha_{\mathrm{B}} \, dV \qquad \ldots\ldots(7.6)$$

where

$$\alpha_{\mathrm{B}} = \sum_{n=2}^{\infty} \alpha_n. \qquad \ldots\ldots(7.7)$$

If the star radiates as a black body

$$\mathcal{N}(h\nu_0) = \pi R_*^2 c \int_{\nu_*}^{\infty} \frac{\rho_\nu(T_*)}{h\nu}\, d\nu \qquad \ldots\ldots(7.8)$$

where R_* is the star radius and

$$\rho_\nu(T) = 8\pi h\nu^3/c^3[\exp(h\nu/kT) - 1].$$

We then have

$$\mathcal{N}(h\nu_0) = \left(\frac{8\pi^2 I_H^3}{h^3 c^2}\right) R_*^2 \frac{F_1(h\nu_0 y/I_H)}{y^3} \qquad \ldots\ldots(7.9)$$

where

$$y = I_H/kT_* \qquad \ldots\ldots(7.10)$$

and where

$$F_1(x) = \int_x^{\infty} \frac{z^2}{e^z - 1}\, dz. \qquad \ldots\ldots(7.11)$$

The function $F_1(x)$ is tabulated by Zanstra (1931).

Assuming constant density within the radius $R(H^+)$ of the H^+ region we obtain

$$\frac{4\pi}{3} R^3(H^+)\alpha_B N_e N(H) = \left(\frac{8\pi^2 I_H^3}{h^3 c^2}\right) R_*^2 \frac{F_1(y)}{y^3} \qquad \ldots\ldots(7.12)$$

where $N(H)$ is the total H density (not to be confused with the density $N(H^0)$ of neutral H).

With $\alpha_B = 2{\cdot}58 \times 10^{-13}$ cm^3 sec^{-1} for $T_* = 10^4$ deg K, we have†

$$R(H^+) = 2{\cdot}41 \times 10^{20}\{N_e N(H)\}^{-1/3}(R_*/R_\odot)^{2/3} \frac{F_1^{1/3}(y)}{y} \text{ cm} \qquad \ldots\ldots(7.13)$$

where $R_\odot$ is the solar radius ($6{\cdot}951 \times 10^{10}$ cm).

For a pure He nebula we obtain a He^{+2} region inside a He^+ region. Using

$$I(He^+) = 4I_H, \quad I(He^0) = 1{\cdot}807 I_H, \quad \alpha_B(He^+) = 1{\cdot}54 \times 10^{-12} \text{ cm}^3 \text{ sec}^{-1}$$

and $\alpha_B(He^0) \simeq \alpha_B(H^0)$ we have

$$R(He^{+2}) = 1{\cdot}33 \times 10^{20}\{N_e N(He)\}^{-1/3}(R_*/R_\odot)^{2/3} \frac{F_1^{1/3}(4y)}{y} \text{ cm} \qquad \ldots\ldots(7.14)$$

and

$$R(He^+) = 2{\cdot}41 \times 10^{20}\{N_e N(He)\}^{-1/3}(R_*/R_\odot)^{2/3} \frac{F_1^{1/3}(1{\cdot}807y)}{y} \text{ cm}. \qquad \ldots\ldots(7.15)$$

In writing (7.15) we have assumed $R^3(He^+) \gg R^3(He^{+2})$; this is a good approximation for all T_*. It should be noted that the radiation absorbed in ionizing He^+ is still available for ionizing He^0 since each recombination to He^+ eventually produces a He II Lyα photon which may ionize He^0 (allowance for He^+ $2s \rightarrow 1s$ two-photon emission would introduce only a minor modification since most $2s \rightarrow 1s$ transitions give one photon capable of ionizing He^0).

Figure 5 shows the radii as functions of T_* calculated for $R_* = R_\odot$, $N(H) = 10^4$ cm^{-3} and $N(He) = 2 \times 10^3$ cm^{-3}. It is seen that $R(H^+)$ is of order 10^{17} cm for $T_* \simeq 4 \times 10^4$ deg K.

We consider what happens with a mixture of H and He. The outer radius of the ionized region is always determined by $R(H^+)$ calculated from (7.13). If the

† Our treatment is more accurate than that of Strömgren but our final results are greater by only 0·1 in log $R(H^+)$.

He abundance is very small and the temperature very high there is a single ionized region within which all the He is doubly ionized. In this case the calculated radii $R(He^{+2})$ and $R(He^{+})$ will be greater than $R(H^{+})$ but these will not have any physical meaning. An intermediate case, which we shall refer to as that of incomplete absorption by He^{0}, corresponds to calculated radii such that

$$R(He^{+2}) < R(H^{+}) < R(He^{+}).$$

We then have He^{+} and H^{+} ions for $R(He^{+2}) < R < R(H^{+})$ and He^{+2} and H^{+} for $R < R(He^{+2})$. In the third case, that of complete absorption by He^{0}, we obtain $R(He^{+2}) < R(He^{+}) < R(H^{+})$. We then have He^{+2} and H^{+} ions for $R < R(He^{+2})$, He^{+} and H^{+} for $R(He^{+2}) < R < R(He^{+})$ and He^{0} and H^{+} for $R(He^{+}) < R < R(H^{+})$.

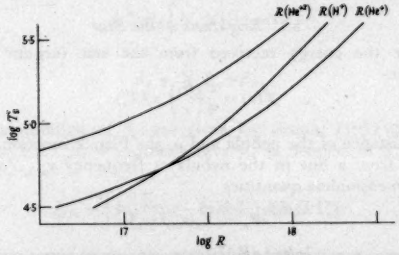


Figure 5. Radii of ionized regions (R in cm, T_* in °K). Calculations for $R_* = R_\odot$, $N(H) = 10^4$ cm^{-3}, $N(He) = 2 \times 10^3$ cm^{-3}.

It may be noted that the radii can be calculated neglecting the mixture of elements since absorption by H^{0} will be small in the He^{+2} and He^{+} regions and since most recombinations to He^{+} and He^{0} produce quanta capable of ionizing H^{0}.

7.4. *Determination of He/H Abundance Ratios*

In §2.7 we determined the ratios $N(He^{+2})/N(H^{+})$ and $N(He^{+})/N(H^{+})$; these are the ratios of the total amount of He^{+2} or He^{+} to the total amount of H^{+}. If the absorption by He^{0} is incomplete there is no He^{0} in the ionized region and the chemical abundance ratio is equal to $N(He^{+2})/N(H^{+})$ plus $N(He^{+})/N(H^{+})$. In §8.4 it will be shown that there is certainly incomplete He^{0} absorption if $I(4686)/I(5876)$ is greater than 1·2, which is the case for many planetaries.

For the case of complete absorption by He^{0} the ionic abundance ratios are determined by the numbers of ionizing quanta and one cannot determine† the chemical abundance ratio without additional observations of the apparent sizes of the nebula in He I and H I radiations. Mathis (1957 c) reports that He I and H I images are of comparable size for a number of planetaries but there may be objects for which there is incomplete absorption by He^{0}.

† Assuming complete absorption by He^{0} we may calculate a temperature $T_*(He\,I/H\,I)$ from the relative H I and He I line intensities (§ 8). If we now calculate the radii using abundance ratios obtained neglecting He^{0} we automatically obtain $R(He^{+}) = R(H^{+})$ for $T_* = T_*(He\,I/H\,I)$.

§ 8. Star Temperature

Assuming that the central star radiates as a black body we have two parameters to determine, the star temperature T_* and the star radius R_*. We therefore require observations of two quantities which depend on the stellar radiation field in different ways. The original work of Zanstra (1931) contains the basic ideas of all the methods we shall describe. In most previous work only relative intensities have been considered and the procedure has therefore been to determine T_* on eliminating R_*; since absolute intensities are now available we seek simultaneous solutions for both T_* and R_*. Our treatment also differs in detail from those given before in that we obtain improved accuracy by making use of developments in the theory of recombination spectra.

8.1. *Brightness of the Star*

Let $\mathscr{J}(\nu)$ be the energy received from the star (erg cm^{-2} sec^{-1} per unit frequency); then

$$\mathscr{J}(\nu) = \frac{c}{4}\left(\frac{R_*}{r}\right)^2 \rho_\nu(T_*) \qquad \ldots\ldots(8.1)$$

where r is the distance of the nebula and ρ_ν the Planck function. Let $\mathscr{I}_n$ be the energy received from a line in the nebula at frequency ν_n. Zanstra (1931) has measured the dimensionless quantities

$$\mathscr{A}_n = \mathscr{I}_n/\nu_n \mathscr{J}(\nu_n). \qquad \ldots\ldots(8.2)$$

We obtain

$$\mathscr{I}_n = \frac{2\pi h\nu_n^4}{c^2}\left(\frac{R_*}{r}\right)^2 \mathscr{A}_n\{\exp(h\nu_n/kT_*) - 1\}^{-1}. \qquad \ldots\ldots(8.3)$$

8.2. *Numbers of Ionizing Quanta*

The number of stellar quanta beyond the Lyman limit is

$$\mathscr{N}(I_H) = \left(\frac{8\pi^2 I_H^3}{h^3 c^2}\right) R_*^2 \frac{F_1(y)}{y^3}, \quad (y = I_H/kT_*) \qquad \ldots\ldots(8.4)$$

(§ 7.3). Assuming complete absorption this is equal to the number of recombinations on $n \geqslant 2$, $\alpha_B \int N_e N(H^+) dV$. Using (2.44) for the absolute Hβ intensity we obtain

$$\mathscr{I}(4861) = \left(\frac{2\pi I_H^3}{h^3 c^2}\right) h\nu(4861)\left(\frac{R_*}{r}\right)^2 \frac{\alpha_{42}(H^0)}{\alpha_B(H^0)} \frac{F_1(y)}{y^3}. \qquad \ldots\ldots(8.5)$$

In the same way we obtain

$$\mathscr{I}(5876) = \left(\frac{2\pi I_H^3}{h^3 c^2}\right) h\nu(5876)\left(\frac{R_*}{r}\right)^2 \frac{\alpha_{3\,^1D,\,2\,^1P}(He^0)}{\alpha_B(He^0)} \frac{F_1(1{\cdot}807\,y)}{y^3} \qquad \ldots\ldots(8.6)$$

assuming complete absorption by He^0 and

$$\mathscr{I}(4686) = \left(\frac{2\pi I_H^3}{h^3 c^2}\right) h\nu(4686)\left(\frac{R_*}{r}\right)^2 \frac{\alpha_{43}(He^+)}{\alpha_B(He^+)} \frac{F_1(4y)}{y^3} \qquad \ldots\ldots(8.7)$$

assuming complete absorption by He$^+$.

8.3. *The Thermal Balance*

The problem of the energy balance may be approached from the standpoint of conservation of radiant energy (Aller, Baker and Menzel 1939, Menzel and Aller 1941 c, Aller 1953 a) or of conservation of electron kinetic energy (Spitzer 1948). We adopt the latter approach.

The electrons gain kinetic energy in photo-ionization and lose kinetic energy in free–free transitions, recombination and forbidden line excitation. The kinetic energy gain due to ionization by quanta produced in the nebula is equal to the loss due to ground state recombinations; these two contributions may therefore be neglected in considering the energy balance. The energy loss due to excited state recombinations plus free–free transitions is

$$\boldsymbol{k}T_e\beta_B\int N_e N(H^+)\,dV \qquad \ldots\ldots(8.8)$$

where

$$\beta_B = \sum_{n=2}^{\infty}\beta_n + \beta_{ff}. \qquad \ldots\ldots(8.9)$$

Tables for the calculation of β_n are given by Seaton (1959 a). For free–free transitions we have $\boldsymbol{k}T_e\beta_{ff} = \int \gamma_{ff}\,d\boldsymbol{h}\nu$ where γ_{ff} is defined in § 2.8. This gives $\beta_{ff} = \frac{1}{2}\mathscr{G}Z\lambda^{1/2}\bar{g}_{III}$ where

$$\bar{g}_{III} = \int \bar{g}_{III}\exp(-\boldsymbol{h}\nu/\boldsymbol{k}T_e)\,d(\boldsymbol{h}\nu/\boldsymbol{k}T_e). \qquad \ldots\ldots(8.10)$$

From the first two terms in the asymptotic expansion for $\bar{g}_{III}$ we obtain

$$\bar{g}_{III} = 1 + 1{\cdot}08(Z^2 I_H/\boldsymbol{k}T_e)^{-1/2}.$$

This is not a very good approximation since the asymptotic expansion is poor in the infra-red—indeed the third term in the expansion gives a divergence in the integral for $\bar{g}_{III}$—but the free–free transitions do not make a large contribution to the energy balance and the error will not be serious. Values of β are given in table 18.

Table 18. Coefficients β_1 and $\beta_B = \sum_{n=2}^{\infty}\beta_n + \beta_{ff}$ for the calculation of Rates of Energy Loss (β_n in cm^3 sec^{-1})

T_e (°K)	1×10^4	2×10^4
$10^{14}\beta_1(H^0)$	14·9	9·9
$10^{14}\beta_B(H^0)$	32·2	20·1
$10^{14}\beta_1(He^+)$	62·8	43·7
$10^{14}\beta_B(He^+)$	168·3	103·9

Let $\mathscr{I}(FL)$ be the total energy received in forbidden lines. Using (2.44) to eliminate the integral in (8.8), we find for the total kinetic energy loss for the nebula

$$4\pi r^2\left\{\mathscr{I}(FL) + \boldsymbol{k}T_e\beta_B\frac{\mathscr{I}(4861)}{\boldsymbol{h}\nu(4861)\,\alpha_{42}(H^0)}\right\}. \qquad \ldots\ldots(8.11)$$

The kinetic energy gain due to ionization by stellar quanta is

$$\pi R_s^2 c \int_{I_{H}/h}^{\infty} (h\nu - I_H) \frac{\rho_\nu}{h\nu} d\nu = \frac{8\pi^2 I_H^4}{h^3 c^3} \left[\frac{F_2(y) - yF_1(y)}{y^4}\right] \quad \ldots\ldots(8.12)$$

where

$$F_2(y) = \int_y^{\infty} \frac{z^3}{e^z - 1} dz$$

is tabulated by Zanstra (1931). Equating (8.11) and (8.12) we have

$$\mathscr{I}(\mathrm{FL}) + kT_e\beta_B \frac{\mathscr{I}(4861)}{h\nu(4861)\alpha_{42}(H^0)} = \frac{2\pi I_H^4}{h^3 c^2}\left(\frac{R_s}{r}\right)^2 \left[\frac{F_2(y) - yF_1(y)}{y^4}\right]. \quad \ldots\ldots(8.13)$$

A more exact expression for the thermal balance could be obtained on allowing for He ionization and recombination.

The sum for $\mathscr{I}(\mathrm{FL})$ is dominated by a few strong lines, particularly those of [O III]. It is probable that most of the forbidden line energy is radiated in the observable spectral region.

8.4. *Summary of Theory Relations*

The ratios of recombination coefficients are insensitive to T_e. Using data from tables 2, 8 and 18 we obtain (with $\mathscr{I}$ in erg cm^{-2} sec^{-1}):

Star : $$\mathscr{I}(4861) = 6{\cdot}71 \times 10^{12}\left(\frac{R_s}{r}\right)^2 \frac{\mathscr{A}(4861)}{(e^{0{\cdot}1875y} - 1)} \quad \ldots\ldots(8.14)$$

H I : $$\mathscr{I}(4861) = 1{\cdot}16 \times 10^{14}\left(\frac{R_s}{r}\right)^2 \frac{F_1(y)}{y^3} \quad \ldots\ldots(8.15)$$

He I : $$\mathscr{I}(5876) = 1{\cdot}17 \times 10^{14}\left(\frac{R_s}{r}\right)^2 \frac{F_1(1{\cdot}807y)}{y^3} \quad \ldots\ldots(8.16)$$

He II : $$\mathscr{I}(4686) = 1{\cdot}41 \times 10^{14}\left(\frac{R_s}{r}\right)^2 \frac{F_1(4y)}{y^3} \quad \ldots\ldots(8.17)$$

Energy : $$\mathscr{I}(\mathrm{FL}) + 3{\cdot}95 \times 10^{-4}\, T_e\, \mathscr{I}(4861)$$
$$= 5{\cdot}43 \times 10^{15}\left(\frac{R_s}{r}\right)^2 \left[\frac{F_2(y) - yF_1(y)}{y^4}\right]. \quad \ldots\ldots(8.18)$$

In principle one could use the ratio of any two of these equations for the determination of T_s. Zanstra (1931) used the ratios (Star/H I), (Star/He I), (Star/He II) and (Star/Energy). The ratio (H I/He II) was first used by Ambarzumian (1932 b) and the ratio (H I/Energy) was first used by Stoy (1933). The method of plotting T_s against R_s has the advantage of showing how errors in the observed quantities affect the derived values of T_s. Errors in the distance r and in the absolute intensities, as distinct from relative intensities, will produce errors in the scale of R_s but not in the solutions for T_s.

From the He II/H I intensities we obtain the relative numbers of ionizing quanta :

$$\frac{I(4686)}{I(4861)} = 1{\cdot}22 \frac{\mathscr{N}(4I_H)}{\mathscr{N}(I_H)}. \quad \ldots\ldots(8.19)$$

This has been discussed by Wurm and Singer (1952) and by Hattori, Kawaguchi, Miyamoto and Saigusa (1953). With complete absorption by H^0 the ratio has an upper limit of 1·22, this limit being approached when nearly all ionizing quanta are emitted beyond the He II limit, which implies $T_* \to \infty$ for a Planck distribution. If $I(4686)/I(4861)$ exceeds 1·22 there is incomplete absorption by H^0. Of the 55 planetaries for which intensities have been measured by Aller (1941, 1951) and by Minkowski and Aller (1956) only one, NGC 4361, appears to have a ratio greater than 1·22 but for this object the intensities are uncertain. Of these 55, 14 have $I(4686)/I(4861) \geqslant 0{\cdot}5$ while for 22 the 4686 line is weak and no intensity ratio is measured.

Fairly large values of $I(4686)/I(4861)$ occur in a number of nebulae of low surface brightness observed by Minkowski (1942) and for some of these it is probable that absorption by H^0 is incomplete. It would be of interest to attempt to measure the [O I] intensities in these objects (see § 7.1).

For complete absorption by He^0 we have

$$\frac{I(4686)}{I(5876)} = 1{\cdot}21 \frac{\mathscr{N}(4I_H)}{\mathscr{N}(1{\cdot}807 I_H)}.$$

If the ratio exceeds 1·21, as it does in many planetaries, absorption by He^0 is incomplete and equation (8.16) cannot be used for determinations of T_*. Incomplete absorption by He^0 does not, of course, necessarily imply any limit to the amount of gas available to be ionized, as would incomplete absorption by H^0 (see § 7.3).

8.5. NGC 6572 *and* 7027

Table 19 gives data for the determination of T_* and R_* for NGC 6572 and 7027. NGC 6572 is one of the two nebulae for which $\mathscr{N}$-values were measured by Zanstra (1931). NGC 7027 is unusual in that no central star has been observed.

Table 19. Data for Determination of T_* and R_*

	Reference	NGC 6572	NGC 7027
$\mathscr{N}(4861)$	1	0·19	—
$I(4861)$	2, 3, 4	100	100
$I(4686)$	2, 3, 4	0·9	50
$I(5876)$	2, 3, 4	16·7	12·7
$I(\mathrm{FL})$	2, 3, 4	1700	2100
c		0·74	1·22
r(parsec)	5	1230	2130
$\mathscr{N}(4861)$ erg cm^{-3} sec^{-1}	5	$9{\cdot}3 \times 10^{-10}$	$12{\cdot}8 \times 10^{-10}$
T_*	6	$1{\cdot}2 \times 10^4$	$1{\cdot}7 \times 10^4$

References

1. Zanstra (1931).
2. Aller (1941) (NGC 6572).
3. Mathis (1957 c) (NGC 6572).
4. Aller, Bowen and Minkowski (1955) (NGC 7027).
5. Aller (1956, p. 149).
6. Figure 2.

Note: (i) Intensities have been corrected for reddening using quoted values of c.
(ii) 1 parsec = $3{\cdot}08 \times 10^{18}$ cm.

Baade (see Liller 1955) estimates that the central star of NGC 7027 can be no brighter than the 18th magnitude; this is to be compared with a photographic magnitude of $m = 12$ for the central star of NGC 6572 (Berman 1937). The

[O I] lines indicate that both of these nebulae are optically thick for H absorption. They differ in that NGC 6572 has weak He II lines indicating complete He^0 absorption while NGC 7027 has strong He II lines and incomplete He^0 absorption.

The T_s, R_s curves are given in figure 6. For NGC 6572 the direct measurement of visual star brightness is seen to give an estimate of R_s which is insensitive to the assumed value of T_s. Combining observations of the star and of the H I lines we obtain $T_s = 4{\cdot}2 \times 10^4$ and $R_s = 1{\cdot}5R_\odot$; this temperature agrees closely with that obtained by Zanstra (1931). Essentially similar results are given by the He I lines and by the energy balance but higher temperatures are indicated by He II. With $R_s = 1{\cdot}5R_\odot$ the He II intensity gives $T_s = 5{\cdot}5 \times 10^4$ deg K. The discrepancy is certainly real, the number of He II quanta being 50 times greater than the number for $T_s = 4{\cdot}2 \times 10^4$. This shows that the stellar ultra-violet

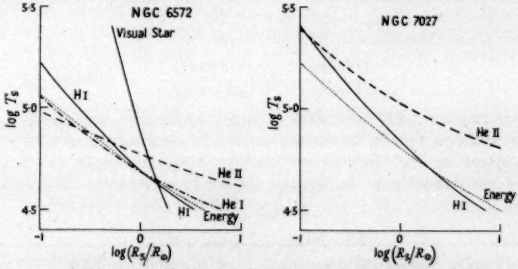


Figure 6. (T_s, R_s) relations for two planetary nebulae.

cannot be represented by a single black-body distribution but for NGC 6572 the number of quanta beyond the He II Lyman limit is still only a small fraction of the total number of ionizing quanta; from the observed intensities we have $\mathcal{N}(4I_H)/\mathcal{N}(I_H) = 0{\cdot}7 \times 10^{-2}$ compared with $\mathcal{N}(4I_H)/\mathcal{N}(I_H) = 1{\cdot}5 \times 10^{-4}$ for $T_s = 4{\cdot}2 \times 10^4$.

From the NGC 7027 curves for H I and the energy balance we obtain $T_s = 5{\cdot}2 \times 10^4$ deg K and $R_s = 1{\cdot}9R_\odot$; an error of 20% in the measured ratio $I(\mathrm{FL})/I(4861)$ would here give an error of $\pm 0{\cdot}1$ in $\log T_s$ and $\pm 0{\cdot}2$ in $\log R_s$. With $R_s = 1{\cdot}9R_\odot$ we obtain $T_s = 9 \times 10^4$ deg K from the He II curve. It is again evident that the stellar ultra-violet cannot be represented by a single black-body distribution and it is now no longer true that only a small fraction of the ionizing quanta are emitted beyond the He II limit; from the observed intensities $\mathcal{N}(4I_H)/\mathcal{N}(I_H) = 0{\cdot}41$.

From figure 6 we see that the He II/H I ratios will give large values of T_s and small values of R_s; from the He II and H I curves for NGC 7027 we obtain $T_s = 2{\cdot}3 \times 10^5$ and $R_s = 0{\cdot}12R_\odot$. Due to the faintness of the central star for NGC 7027 we cannot be sure that the small value of R_s is incorrect but one can reject such values for certain other nebulae. Thus for NGC 7662 the ratio $I(4686)/I(4861)$ is similar to that for NGC 7027 and $\mathcal{I}(4861)$ is smaller by a factor of about 4. From the He II and H I curves we would therefore obtain

$T_* = 2{\cdot}3 \times 10^5$ and $R_* = 0{\cdot}06 R_\odot$. This small radius would be difficult to reconcile with the fact that the central star is not particularly faint ($m = 12{\cdot}7$ according to Berman 1937).

8.6. *Discussion of the Stellar Ultra-violet*

We have considered a low excitation nebula (weak He II lines) and a high excitation nebula (strong He II lines). In both cases we find that the star radiation cannot be described completely in terms of a Planck distribution. For low excitation nebulae it may be that much of the ultra-violet is characterized by a temperature of about 4×10^4 deg K but there is some excess of high energy radiation. For high excitation nebulae one might still consider temperatures of order 4×10^4 deg K for some way beyond the H I Lyman limit but much higher temperatures, of order 10^5 deg K or more, beyond the He II Lyman limit. Such trends are consistent with results of model atmosphere calculations for hot stars (see Aller 1956, Ch. VI) which give a depression of the emergent flux to the high frequency side of the H I Lyman limit. One might also consider the possibility of the central stars having a large amount of high energy coronal type radiation.

Study of the ionization equilibrium for complex ions could give further information about the distribution of stellar ultra-violet. Measurements of the quantities $\mathscr{A}$ (equation (8.2)) for many more nebulae would be very valuable since, combined with absolute brightness measurements, these give good estimates for the star radii.

The observed spectra of some central stars of planetaries are discussed by Aller (1956, Ch. VI).

§9. The Equation of State

The electron pressure is $p_e = N_e \boldsymbol{k} T_e$. Defining the electron specific volume as $v_e = 1/N_e$ we have $p_e v_e = \boldsymbol{k} T_e$. A second relation between p_e, v_e and T_e is obtained from the theory of the thermal balance. Neglecting deactivation processes, the electron temperature depends only on the frequency distribution of radiant energy and on the ionic composition of the nebula. If the ionic composition did not depend on the density the equation of state would be $p_e v_e =$ constant. Zanstra (1955) has considered the equation of state obtained when the density variation of the ionic composition is taken into account and has shown that, for sufficiently high star temperatures, this gives the possibility that two phases of different specific volume may co-exist at the same pressure. This could explain the existence of local high density condensations.

The thermal balance is considered per cm^3 per second for some point inside the nebula rather than for the entire nebula, assuming complete absorption, as in §8.3. The energy loss due to recombination and free–free transitions is

$$\beta \boldsymbol{k} T_e N_e N(\mathrm{H}^+)$$

where

$$\beta = \sum_{n=1}^{\infty} \beta_n + \beta_{\mathrm{ff}}.$$

Let $\mathscr{E}$ be the energy radiated in forbidden lines per cm^3 per second. The equation for the thermal balance is

$$\beta \boldsymbol{k} T_e N_e N(\mathrm{H}^+) + \mathscr{E} = N(\mathrm{H}^0) \int \frac{c \rho_\nu(\mathrm{neb})}{\boldsymbol{h}\nu} (\boldsymbol{h}\nu - I_{\mathrm{H}})\, a_\nu(\mathrm{H}^0)\, d\nu \qquad \ldots\ldots (9.1)$$

where the right-hand side is the energy input due to hydrogen photo-ionization, ρ_ν(neb) being the radiation density in the nebula. The hydrogen ionization equation is

$$\alpha N_e N(H^+) = N(H^0) \int \frac{c\rho_\nu(\text{neb})}{h\nu} a_\nu(H^0)\, d\nu \qquad \ldots\ldots(9.2)$$

where

$$\alpha = \sum_{n=1}^{\infty} \alpha_n.$$

Dividing (9.1) by (9.2) we obtain

$$\phi(T_e)\, kT_e + \frac{\mathscr{E}}{\alpha N_e N(H^+)} = kT_{ph} \qquad \ldots\ldots(9.3)$$

where $\phi = \beta/\alpha$ and where

$$kT_{ph} = \frac{\int \left(\frac{\rho_\nu}{h\nu}\right)(h\nu - I_H)\, a_\nu\, d\nu}{\int \left(\frac{\rho_\nu}{h\nu}\right) a_\nu\, d\nu} \qquad \ldots\ldots(9.4)$$

is the mean kinetic energy of the photo-electrons. If ρ_ν(neb) has the frequency dependence of a Planck function for the temperature T_* we have

$$T_{ph} = \psi(y) T_*, \quad (y = I_H/kT_*). \qquad \ldots\ldots(9.5)$$

The function $\psi(y)$, calculated with $g_{II} = 1$ (§ 2.1), is tabulated by Spitzer (1948) for $y \geqslant 0{\cdot}5$ and is given in table 20 for $y \leqslant 0{\cdot}5$. Using formulae of § 8.3 together with formulae given by Seaton (1959 a) we obtain

$$\phi(T_e) = \frac{0{\cdot}429 + \frac{1}{2}\ln\lambda + 1{\cdot}18\lambda^{-1/3}}{0{\cdot}429 + \frac{1}{2}\ln\lambda + 0{\cdot}47\lambda^{-1/3}}, \quad \lambda = I_H/kT_e. \qquad \ldots\ldots(9.6)$$

Table 20

y	0·00	0·05	0·10	0·15	0·2	0·3	0·4	0·5
$\psi(y)$	0·000	0·116	0·182	0·235	0·280	0·352	0·409	0·450

For many nebulae most of the forbidden line energy is radiated in the N_1, N_2 [O III] lines. Considering this contribution alone, and assuming the density to be too low for deactivation to be important,

$$\mathscr{E} = C_1 N_e N(O^{+2}) \frac{\exp(-\Delta E/kT_e)}{T_e^{1/2}} \qquad \ldots\ldots(9.7)$$

where (§ 4.1)

$$C_1 = 8{\cdot}63 \times 10^{-6} (\Omega_{1,2}/\omega_1)\, \Delta E = 6{\cdot}1 \times 10^{-18} \text{ erg cm}^3 \text{ sec}^{-1},$$

$\Delta E = 2{\cdot}49$ ev being the excitation energy. Substitution in (9.3) gives

$$\phi kT_e + C_1 \frac{\exp(-\Delta E/kT_e)}{\alpha(H^0)\, T_e^{1/2}} \frac{N(O^{+2})}{N(H^+)} = kT_{ph}. \qquad \ldots\ldots(9.8)$$

Figure 7 gives T_e as a function of T_*, calculated from (9.5) and (9.8), for various values of $N(O^{+2})/N(H^+)$. In most planetaries $N(O^{+2})/N(H^+) \sim 10^{-4}$. With $T_* \sim 5 \times 10^4$ this gives $T_e \sim 1{\cdot}5 \times 10^4$ deg K.

We now consider how $N(O^{+2})/N(H^+)$ varies with N_e, T_e. For NGC 7027 it has been shown that O^{+3} is the most abundant oxygen ion and that $N(O^{+3})/N(H^+)$ is approximately equal to the chemical abundance ratio $N(O)/N(H)$. A similar position may be expected for other high excitation nebulae. Using the equation for ionization equilibrium between O^{+2} and O^{+3} we obtain

$$\phi k T_e + C_2 \frac{\exp(-\Delta E/kT_e)}{T_e^{1/2}} N_e = k T_{ph} \qquad \ldots\ldots(9.9)$$

where

$$C_2 = \frac{\alpha(O^{+2})}{\alpha(H^0)} \frac{N(O^{+3})}{N(H^+)} \frac{C_1}{\int c\{\rho_\nu(\text{neb})/h\nu\} a_\nu(O^{+2})\, d\nu}. \qquad \ldots\ldots(9.10)$$

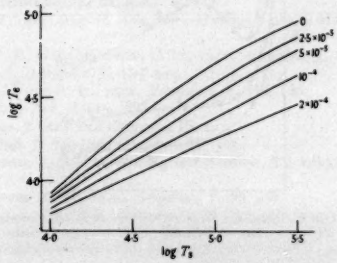


Figure 7. (log T_e, log T_s) for values of $N(O^{+2})/N(H^+)$ ranging from 0 to 2×10^{-4}.

The ratio $\alpha(O^{+2})/\alpha(H^0)$ will vary slowly with T_e, and C_2 may therefore be treated as a constant if O^{+3} has an abundance much greater than that of all other oxygen ions. From $p_e = N_e k T_e$ we obtain

$$p_e = \frac{k^2}{C_2} \exp(\Delta E/kT_e)\, T_e^{3/2} (T_{ph} - \phi T_e). \qquad \ldots\ldots(9.11)$$

This has the form of the equation obtained by Zanstra (1955). For a given value of $T_{ph} = \phi T_s$ we may use (9.11) to calculate p_e as a function of T_e and then obtain v_e from $v_e = kT_e/p_e$. Figure 8 shows (p_e, v_e) plots for three values of T_s.

Above a certain critical temperature, $T_{ph} = 1{\cdot}1\times10^5$ deg K corresponding to $T_s = 2{\cdot}1\times10^5$ deg K, the equation of state gives (p_e, v_e) curves with maxima and minima. Between the maximum and minimum points the equilibrium is unstable since compression leads to a reduction in pressure. For a given pressure there is the possibility of obtaining two positions of stable equilibrium, such as those marked A and B in figure 8, and hence the co-existence of two phases of different specific volume.

Observational evidence for the existence of local high density condensations is obtained from direct photographs of nebulae, which often show an irregular mottled structure. Further, it has been seen in §3.3 that the absolute brightness and the relative intensities often show T_e, N_e to be non-uniform. For some nebulae the relative [O II] intensities are alone sufficient to indicate the existence

of non-uniformities. In order to explain the existence of condensations in terms of the (p_e, v_e) curves of figure 8, very high values of T_{ph} must be assumed. In many cases these correspond to values of T_s considerably in excess of those obtained by the methods of § 8. Some reduction in the critical temperature for phase separation might be obtained in a more complete theory which takes more stages of ionization into account and allows for cooling by other ions, particularly those

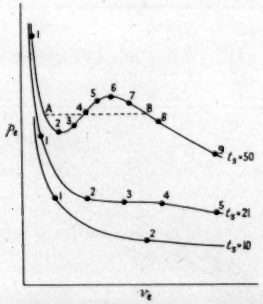


Figure 8. Plots of p_e against v_e calculated from (9.11) and (9.5) for various values of $t_s = 10^{-4} T_s$. The curve for $t_s = 21$ has a point of inflection of zero slope and the curves for $t_s > 21$ have maxima and minima. The electron temperature varies along each (p_e, v_e) curve; some values of $t_e = 10^{-4} T_e$ are marked on the curves.

of nitrogen, but it seems to be improbable that the equation of state can explain non-uniformities observed in all cases.

NGC 7293 is a planetary of particular interest. Photographs by Baade and by Minkowski (see Zanstra 1955) show curious comet-like condensations. For this nebula Zanstra obtains $T_s = 8{\cdot}5 \times 10^4$ deg K assuming complete absorption but since this object has an open structure (Zanstra considers it to be three turns of a helix) the star temperature must in fact be much greater and may well exceed the critical value of $2{\cdot}1 \times 10^5$ deg K required to explain the existence of condensations in terms of equation (9.11).

References

Aller, L. H., 1941, *Astrophys. J.*, **93**, 236; 1951, *Ibid.*, **113**, 125; 1953 a, *Ibid.*, **118**, 547; 1953 b, *Astrophysics*, Vol. 2, p. 327 (New York: Ronald Press); 1954, *Astrophys. J.*, **120**, 401; 1956, *Gaseous Nebulae* (London: Chapman and Hall); 1957, *Astrophys. J.*, **125**, 84.

Aller, L. H., Baker, J. G., and Menzel, D. H., 1939, *Astrophys. J.*, **89**, 587, **90**, 271.

Aller, L. H., Bowen, I. S., and Minkowski, R., 1955, *Astrophys. J.*, **122**, 62.

Aller, L. H., and Menzel, D. H., 1945, *Astrophys. J.*, **102**, 239.

Aller, L. H., and Minkowski, R., 1956, *Astrophys. J.*, **124**, 110.

Ambarzumian, V. A., 1932 a, *Mon. Not. R. Astr. Soc.*, **93**, 50; 1932 b, *Poulkovo Obs. Circular*, **8**, No. 4.

Baker, J. G., and Menzel, D. H., 1938, *Astrophys. J.*, **88**, 52.

Barbier, D., and Andrillat, H., 1954, *C.R. Acad. Sci., Paris*, **238**, 1099.

Bates, D. R., and Damgaard, A., 1949, *Phil. Trans. Roy. Soc.* A, **242**, 101.

BERMAN, L., 1936, *Mon. Not. R. Astr. Soc.*, **96**, 890 ; 1937, *Lick Obs. Bull.*, **18**, 73.
BOHM, D., and ALLER, L. H., 1947, *Astrophys. J.*, **105**, 131.
BOWEN, I. S., 1928, *Astrophys. J.*, **67**, 1 ; 1934, *Publ. Astr. Soc. Pacific*, **46**, 146.
BOWEN, I. S., and WYSE, A. B., 1939, *Lick Obs. Bull.*, **19**, 1.
BURGESS, A., 1958, *Mon. Not. R. Astr. Soc.*, **118**, 477.
BURGESS, A., and SEATON, M. J., 1960 a, *Mon. Not. R. Astr. Soc.*, **120**, No. 2 ; 1960 b, *Ibid.*, **120**, in the press.
CAYREL, R., 1957, *Thèses*, Université de Paris.
CHAMBERLAIN, J. W., 1953, *Astrophys. J.*, **117**, 387.
CHANDRASEKHAR, S., 1935, *Z. f. Astrophys.*, **9**, 266 ; 1945, *Rev. Mod. Phys.*, **17**, 138, *Astrophys. J.*, **102**, 402.
DIVAN, L., 1954, *Thèses*, Université de Paris.
DUFAY, J., 1957, *Galactic Nebulae and Interstellar Matter* (London : Hutchinson).
DUNCAN, J. C., 1937, *Astrophys. J.*, **86**, 496.
GARSTANG, R. H., 1951, *Mon. Not. R. Astr. Soc.*, **111**, 115 ; 1957, *Mém. Soc. Roy. Sci. Liège*, **20**, 497.
GREEN, L. C., RUSH, P. P., and CHANDLER, C. D., 1957, *Astrophys. J. Suppl. Series*, **3**, 37.
GREENSTEIN, J. L., 1946, *Astrophys. J.*, **104**, 414.
GREENSTEIN, J. L., and PAGE, T. L., 1951, *Astrophys. J.*, **114**, 106.
GURZADIAN, G. A., 1952, *Astr. Zhurn.*, **29**, 111.
HAGIHARA, Y., 1938, *Jap. J. Astr. and Geophys.*, **15**, 1.
HATANAKA, T., 1944, *Jap. J. Astr. and Geophys.*, **20**, 505.
HATTORI, A., KAWAGUCHI, I., MIYAMOTO, S., and SAIGUSA, T., 1953, *Publ. Astr. Soc. Japan*, **4**, 152.
HEBB, M. H., and MENZEL, D. H., 1940, *Astrophys. J.*, **92**, 408.
DE JONG, J. H., 1951, *Bull. Astr. Inst. Netherlands*, **11** (No. 428), 345.
KIPPER, A. Y., 1952, *Tartu Astr. Obs. Publ.*, **32**, 63.
KOELBLOED, D., 1956, *Bull. Astr. Inst. Netherlands*, **12** (No. 465), 341.
LILLER, W., 1955, *Astrophys. J.*, **122**, 240.
LILLER, W., and ALLER, L. H., 1954, *Astrophys. J.*, **120**, 48.
MATHIS, J. S., 1957 a, *Astrophys. J.*, **125**, 318 ; 1957 b, *Ibid.*, **125**, 328 ; 1957 c, *Ibid.*, **126**, 493.
MENZEL, D. H., and ALLER, L. H., 1941 a, *Astrophys. J.*, **94**, 436 ; 1941 b, *Ibid.*, **93**, 195 ; 1941 c, *Ibid.*, **94**, 30.
MENZEL, D. H., and PEKERIS, C. L., 1935, *Mon. Not. R. Astr. Soc.*, **96**, 77.
MINKOWSKI, R., 1942, *Astrophys. J.*, **95**, 243 ; 1953, *Publ. Astr. Soc. Pacific*, **65**, 161.
MINKOWSKI, R., and ALLER, L. H., 1956, *Astrophys. J.*, **124**, 93.
MIYAMOTO, S., 1950, *Publ. Astr. Soc. Jap.*, **2**, 23.
NIKITIN, A. A., 1950, *Uchenye Zapiski Leningradskogo Gosudarstvennogo Universiteta*, No. 136, p. 47.
PAGE, T. L., 1936, *Mon. Not. R. Astr. Soc.*, **96**, 604 ; 1942, *Astrophys. J.*, **96**, 78.
PAGE, T. L., and GREENSTEIN, J. L., 1951, *Astrophys. J.*, **114**, 98.
PLASKETT, H. H., 1928, *Publ. Dom. Astr. Obs.* (*Victoria*), **4**, 187.
POTTASCH, S. R., 1958, *Thesis*, University of Colorado.
SAIGUSA, T., 1951, *Publ. Astr. Soc. Jap.*, **4**, 17.
SEARLE, L., 1958, *Astrophys. J.*, **128**, 489.
SEATON, M. J., 1953, *Proc. Roy. Soc.* A, **218**, 400 ; 1954 a, *Ann. d'Astrophys.*, **17**, 550 ; 1954 b, *Mon. Not. R. Astr. Soc.*, **114**, 154 ; 1954 c, *Mém. Soc. Roy. Sci. Liège*, **15**, 462 ; 1955, *Mon. Not. R. Astr. Soc.*, **115**, 279 ; 1958, *Rev. Mod. Phys.*, **30**, 979 ; 1959 a, *Mon. Not. R. Astr. Soc.*, **119**, 81 ; 1959 b, *Ibid.*, **119**, 90 ; 1960, *Ibid.*, **120**, in the press.
SEATON, M. J., and OSTERBROCK, D. E., 1957, *Astrophys. J.*, **125**, 66.
SHAÏN, G. A., 1934, *Tsirkulyavy Glavnoĭ Astronomicheskoĭ Observatorii*, No. 11, p. 8.
SOBOLEV, V. V., 1944, *Astr. Zhurn.*, **21**, 143 ; 1955, *Vestnik Leningradskogo Gosudarstvennogo Universiteta*, **5**, 85.
SPITZER, L., 1948, *Astrophys. J.*, **107**, 6 ; 1956, *Physics of Fully Ionized Gases* (New York : Interscience).
SPITZER, L., and GREENSTEIN, J. L., 1951, *Astrophys. J.*, **114**, 407.

STOY, R. H., 1933, *Mon. Not. R. Astr. Soc.*, **93**, 588.
STRÖMGREN, B., 1939, *Astrophys. J.*, **89**, 526.
TRAVING, G., 1955, *Z. Astrophys.*, **36**, 1 ; 1957, *Ibid.*, **41**, 215.
UNNO, W., 1950, *Publ. Astr. Soc. Japan*, **2**, 53; 1952 a, *Ibid.*, **3**, 158; 1952 b, *Ibid.*, **4**, 100; 1952 c, *Ibid.*, **3**, 178; 1955, *Ibid.*, **7**, 81.
VORONTSOV-VELYAMINOV, B. A., 1948, *Gaseous Nebulae and New Stars* (Moscow : U.S.S.R. Academy of Sciences).
VAN DE HULST, H. C., 1952, *Astrophys. J.*, **115**, 331.
WHITFORD, A. E., 1948, *Astrophys. J.*, **107**, 102 ; 1958, *Astron. J.*, **63**, 201.
WILSON, O. C., 1958, *Rev. Mod. Phys.*, **30**, 1025.
WILSON, O. C., and ALLER, L. H., 1951, *Astrophys. J.*, **114**, 421.
WURM, K., 1951, *Die planetarischen Nebel* (Berlin : Akademie-Verlag) ; 1954, *Mém. Soc. Roy. Sci. Liège*, **14**, 387.
WURM, K., and SINGER, O., 1952, *Z. Astrophys.*, **30**, 387.
WYSE, A. B., 1942, *Astrophys. J.*, **95**, 356.
YADA, B., 1953, *Publ. Astr. Soc. Jap.*, **5**, 128 ; 1954, *Ibid.*, **6**, 76.
YADA, B., and ŌSAKI, T., 1957, *Publ. Astr. Soc. Jap.*, **9**, 82.
ZANSTRA, H., 1927, *Astrophys. J.*, **65**, 50 ; 1931, *Publ. Dom. Astr. Obs.* (*Victoria*), **4**, 209 ; 1934, *Mon. Not. R. Astr. Soc.*, **95**, 84 ; 1949, *Bull. Astr. Inst. Netherlands*, **9** (No. 401), 1 ; 1951, *Ibid.*, **11** (No. 429), 359 ; 1955, *Vistas in Astronomy* (ed. A. Beer), Vol. 1, p. 256 (London: Pergamon).

BAND STRUCTURE CALCULATIONS IN SOLIDS

By L. PINCHERLE
Bedford College, University of London

CONTENTS

	PAGE
§ 1. Introduction	355
§ 2. Direct experimental evidence	357
§ 3. General review and classification	359
§ 4. Details of the various methods	363
4.1. The cellular method	363
4.2. Methods based on atomic orbitals	366
4.2.1. Tight binding	366
4.2.2. Equivalent orbitals	368
4.2.3. Orthogonalized atomic orbitals	369
4.2.4. Methods based on the use of Wannier functions	369
4.2.5. Semi-localized crystal orbitals	371
4.3. Plane wave methods	371
4.3.1. Simple plane waves	371
4.3.2. Orthogonalized plane waves	375
4.3.3. Modified plane waves	377
4.4. Methods using a constant potential outside atomic spheres	378
4.4.1. Augmented plane waves	378
4.4.2. Scattering matrix method	380
4.4.3. Integral equation approach	381
4.4.4. Use of a constant potential outside atomic spheres in conjunction with the cellular procedure	382
4.5. Methods using empirical parameters	383
4.5.1. The quantum defect method	383
4.5.2. The $\mathbf{k}.\mathbf{p}$ approximation	384
4.5.3. Pseudo-potentials	385
§ 5. The construction of the crystal potential	385
5.1. The approximations most widely used	389
§ 6. Concluding remarks	391
References	392

Abstract. After a brief review of the main experimental evidence on the electronic band structure of solids, the various methods for calculating band structures, either from first principles or from empiric data, are classified and critically examined. Finally the problem of the construction of the crystal potential is discussed in an elementary way.

§ 1. Introduction

In the approximation in which one treats electrons in perfect crystals by considering the behaviour of one particle in a fixed periodic field, one uses generally a scheme (reduced zone scheme) in which the energy E is a multi-valued continuous function of three quantum numbers k_x, k_y, k_z, which in practice may be taken to vary continuously. These form a vector $\mathbf{k}$, which, as far as external forces are concerned, behaves as the momentum of the electron. The values of $\mathbf{k}$ are restricted to a polyhedral domain, the first Brillouin zone (often called improperly just ' the Brillouin zone '). The totality of the energy values corresponding to each branch of the $E(\mathbf{k})$ function constitutes an energy band.

The detailed dependence on $\mathbf{k}$ of the multivalued $E(\mathbf{k})$ function (dispersion relation) is what one calls the electronic band structure of the crystal. For a general exposition of band theory and the early bibliography on band structure calculations see Slater (1956). For a short introduction to band theory see Herman (1955 a). As is well known, $\hbar^{-1}\,\mathrm{grad}_{\mathbf{k}}E(\mathbf{k})$ gives the mean velocity of an electron of momentum $\hbar\mathbf{k}$ and energy E; and the tensor $\hbar^{-2}(\partial^2 E/\partial k_i\,\partial k_j)$ is the reciprocal of the effective mass tensor.

At absolute zero the energy levels are occupied up to a value $E = E_F$ and empty above. E_F is called the Fermi level. In insulators and semiconductors it falls in a gap between two bands; in metals it is within one or more bands and the surface (in momentum space, or $\mathbf{k}$-space) $E(\mathbf{k}) = E_F$ is called the Fermi surface. It may be a closed or an open surface, may consist of separate parts and may have any shape consistent with the symmetry of the crystal.

The $E(\mathbf{k})$ relation for the conduction band or bands of a metal, particularly near the Fermi surface, and for the valence and conduction bands of a semiconductor, is extremely important in determining the electrical, optical, magnetic and thermal properties of the substance. It is thus extremely desirable to be able to derive accurately such a relation from first principles. Unfortunately these calculations are very complicated and no completely satisfactory calculation has been carried out so far. However, those parts of the band structure, mentioned above, which are important for the physical properties, can be determined in detail by a judicious combination of fundamental calculations with various experimental results, and this has been done for several substances. Indeed, in the last few years the emphasis has shifted towards this kind of semi-empirical approach, for example with the so-called $\mathbf{k}.\mathbf{p}$ approximation (Kane 1959) and with Phillips' (1959) pseudo-potentials. Once the $E(\mathbf{k})$ relation has been determined, one can derive the density of states dn_s/dE (number of states per unit volume per unit energy interval), given by

$$\frac{dn_s}{dE} = \frac{2}{8\pi^3}\int \frac{d\sigma}{|\mathrm{grad}_{\mathbf{k}}E|}$$

the integral extending over the constant energy surface $E(\mathbf{k}) = E$ in $\mathbf{k}$-space.

In this article complete validity of the concepts of band theory will be assumed. It is of course an approximation, but it has proved perfectly adequate for discussing the great majority of metal and semiconductor phenomena, down to the finest details. Obviously it must fail to account for interactions between carriers, or between carriers and lattice. Slater (1959) has recently discussed the bases of the theory, and Kohn (1957, 1958, 1959) has shown by field theory methods how from the many-electron Schrödinger equation it follows strictly that the carriers in a semiconductor may be regarded much like free electrons or positrons, provided the free electron mass is replaced by an effective mass (generally a tensor) and the (weak) interaction with any charge in the lattice is taken as a Coulomb one modified by the static dielectric constant of the substance. The proof has not yet been extended to the case when a magnetic field is present. A discussion of the definition of the Fermi surface in metals has been given by Mott (1956).

Within the band approximation the problem of calculating the band structure of a substance from first principles divides into two parts: the setting up of the best one-electron Hamiltonian $\mathscr{H}$, and the solution of the corresponding Schrödinger

equation with the appropriate boundary conditions. The first part has been tackled almost universally from the self-consistent field point of view, but has not received yet any satisfactory treatment, and in fact no self-consistent field in a crystal has been worked out : the nearest thing to a completely self-consistent field is that worked out for aluminium by Heine (1957). The difficulty is that the computation is exceedingly laborious, and is only now becoming possible with the coming into use of large and fast computers : one must in fact find the wave functions, not only the energy, for a very large number of values of **k**, and add up the corresponding charge distributions to obtain the potential for the next stage. To add to the complication, this potential is itself a function of **k**.

The second part of the problem has however attracted much attention and a great variety of methods, some quite ingenious, has been devised for obtaining the band structure, given the type of crystal and the potential. Some of these methods have been pushed to a high degree of approximation, but mainly to test their potentialities rather than to obtain really accurate results, in view of the uncertainty in the potential employed. The very important question of how much the band structure is dependent on the details of the assumed potential has not been elucidated yet, and has only recently attracted some attention. The great similarities of the band structures of crystals with similar structures, e.g. silicon, germanium and the III–V compounds, gives a clear indication that in a proper calculation the details of the potential must not turn out to be of such great importance as some old results seemed to show. Indeed, the order of the terms is generally that given by the simplest free electron picture (§ 4.3.1), and many other qualitative features of the energy bands of metals and even of semiconductors can be understood on the basis of such a picture. The converse of this problem, that is the determination of the potential from a given band structure, has not been fully examined. Steinmann (1957) showed on a one-dimensional model that there are infinite potentials which lead to the same dependence on the energy of the effective mass.

This review will be concerned almost exclusively with the methods employed in band structure calculations, since the main results achieved have been extensively discussed in recent articles (Raynor 1952, Herman 1958, Callaway 1958 a). It may be said to bring up to date an earlier review by Reitz (1955). It will consider mainly the physical principles involved rather than the detailed mathematical techniques. As in the analogous problem of the calculation of molecular properties (Hall 1959), the calculations may be divided into those that try to solve the Schrödinger equation (fundamental methods—the main concern of this article), those that make use of empirical parameters, e.g. Phillips' method of pseudo-potentials, and those using models, e.g. the free electron model.

§ 2. Direct Experimental Evidence

For a long time, students of solid state phenomena were content with a theory which could explain qualitatively the experimental results. It has been only in recent years that, with the help of improved experimental techniques and following the discovery of new phenomena, one has begun to try to obtain complete correlation between theory and experiment.

For this purpose, the details of the band structure have become all important. An example familiar to experimental physicists is that of the electrical conductivity

of a semiconductor. In the early theories it seemed adequate to consider spherical energy surfaces with $E(\mathbf{k})-E(0)$ proportional to k^2. One then finds that, when in a semiconductor the carriers are scattered by acoustical phonons, their mobility should vary with temperature according to a $T^{-3/2}$ law. But deviations from this law were observed in most semiconductors, and one cause of them is the scattering from one energy minimum in k-space to an equivalent one (intervalley scattering), due to a peculiarity of the band structure. Now the fine details of practically all phenomena (magnetoresistance, ratio between drift and Hall mobility, thermoelectric power, optical absorption spectrum, etc.) are influenced by the band structure and, conversely, their study can give at least indirect information on the band structure; in particular, from magnetoresistance studies, one can determine the effective mass tensor. But there are some effects which give direct and detailed information. For reviews see Lax (1958), mainly on semiconductors, and Chambers (1956) on metals. The relation of the shape of the Fermi surface with optical constants and heat capacity has been discussed by Cohen (1958); that with the magnetoresistance (for alkali metals) by García Moliner (1958 a); and that with the thermoelectric power (for noble metals) by Jones (1955).

For metals the most powerful method to-day appears to be the de Haas–van Alphen effect, that is the fact that at low temperature, when the mean free path is large so that the carriers move in quantized orbits under the influence of an applied magnetic field H, the diamagnetic susceptibility is a periodic function of H^{-1}. The period gives directly the extremal cross section of the Fermi surface in planes perpendicular to $\mathbf{H}$. How useful the de Haas–van Alphen effect can be is well demonstrated for instance by Gold's (1958) determination of the Fermi surface of Pb and by the analysis of the band structure of graphite carried out by McClure (1957) (see also Soule and McClure 1959). Similar information is provided by the oscillatory Hall effect and oscillatory magnetoresistance. Cyclotron resonance (e.g. Dresselhaus, Kip and Kittel 1955, on Ge), when feasible, is very powerful and gives very direct information on the effective mass tensor, that is the structure of the band edges, in semiconductors. The quantum theory of cyclotron resonance was given by Luttinger (1956). For metals the equivalent tool is the study of the resonance under conditions of anomalous skin effect. When the skin depth is small compared to the mean free path of the carriers, the only electrons that can be affected by the high frequency field are those moving parallel to the surface, that is situated in small regions of the Fermi surface S_F. Using single crystals, one can thus obtain information on the curvature of this surface at various points. Application of a magnetic field then gives also $(\partial S_F/\partial E)_{extr}$, the rate at which the extremal cross section of the Fermi surface varies with energy, that is the electron velocities at the Fermi surface (Azbel' 1958, Azbel' and Kaner 1958 and references). For an application of the method see, for instance, Aubrey and Chambers' (1957) study of the Fermi surface of bismuth.

Detailed measurements of the optical absorption edge under high resolution (Macfarlane and Roberts 1955) also provide information on the location of the band extrema in k-space. In certain substances (InSb) where the density of states is small, information is obtained on the energy surfaces for an appreciable energy interval. More recently, the oscillations of the absorption coefficient near the edge under application of a magnetic field have provided a new powerful tool (see, for

instance, Roth, Lax and Zwerdling 1959). Also these oscillations are due to the quantization of the electron orbits in the magnetic field.

Also the asymmetry of hot electrons, that is the fact that electric conduction becomes anisotropic in cubic crystals under high electric fields when Ohm's law fails (Shibuya 1955, Sasaki *et al.* 1959), has been suggested as a rapid and easy method for obtaining at least qualitative information on band structure. This effect occurs if the electric field makes different angles with the axes of the various equivalent pockets (valleys) of carriers. Then in some pockets the carriers become ' hotter ' than in others.

In principle, much information on the electron structure of metals, semiconductors and insulators alike, can be provided by x-ray spectroscopy. The intensity distribution in the lines of the emission spectra, when an electron jumps from an occupied state in the valence band on to a vacancy in an inner shell, should give a direct picture of the density of states in the valence band, or at least of the density of states multiplied by the transition probability to the vacant inner level (if this inner level is, for instance, an s-level, the transition probability is highest for states of the valence band having predominantly p character). Similarly, the absorption spectra, when an inner electron jumps to an unoccupied state, should give a picture of the density of states in the empty bands.

However, apart from experimental difficulties, line broadening, presence of satellite lines, and other causes of error, one must remember that, when an electron is removed from an inner shell of a lattice atom of atomic number Z, the atom itself behaves in many respects as an impurity of atomic charge $Z+1$. Impurity states are then set up, and these are what one mainly observes. The separation of the features due to the normal bands and those due to the impurity states requires very careful interpretation of the experimental results. An important factor is the relative lifetimes of the various processes involved. A very fair discussion of the subject has been given by Parratt (1959), who first pointed out the fundamental difficulty due to the local states : for an earlier review see Skinner (1939). While at present little more than the width of the bands is measured reliably, progress has been rapid and it is probable that many details will in future be obtained by this method.

§ 3. General Review and Classification

The first step to decide, when undertaking a band structure calculation, is clearly which type of one-electron wave function one is to employ, and two extreme views are immediately apparent. On one hand one may consider the electrons to be practically free, and thus represented by a single plane wave $\psi_{\mathbf{k}}(\mathbf{r}) = A \exp(i\mathbf{k}.\mathbf{r})$ (which may be normalized, for instance, per unit volume, or per unit crystal cell). Since all reciprocal lattice points are equivalent, to each wave of wave vector $\mathbf{k}$ there corresponds an infinite number of energy values of the form

$$E_n = |\mathbf{k} + \mathbf{K}_n|^2 \dagger \qquad \ldots\ldots(1)$$

where $\mathbf{K}_n$ is any translation which leaves invariant the reciprocal lattice. This will be an appropriate starting point to deal with electrons in metals. It is an adequate

† $\mathbf{k}$ is generally measured in reciprocal Bohr radii, and E in rydberg units.

representation for high energy electrons in any substance. On the other hand, for inner core electrons, whose atomic energy levels are only slightly broadened and displaced by the fact that the atoms form part of a crystal, a description by means of atomic orbitals is most convenient. Also in this case, however, the wave functions are taken to be of the type of molecular orbitals, that is they extend through the whole crystal. Only a few calculations have been made from the point of view of the valence bond model, where electrons are attached to determined atoms.

If it were possible to push the calculation to all degrees of approximation, all starting points would lead to the same result; but for practical reasons one must stop far short of such an ideal, and strive to achieve reasonable results with the least labour. Hence it is essential to start with the approximation most suited to the problem under study.

The next step is to construct those linear combinations of the chosen basic functions which are appropriate to the symmetry of the crystal and the value of $\mathbf{k}$. Having decided this value of $\mathbf{k}$ for which one wants to calculate the energy, there will be a group of operations (rotations and reflections in $\mathbf{k}$-space) which leave $\mathbf{k}$ invariant or transform it into an equivalent point $\mathbf{k}+\mathbf{K}_n$: the correct linear combination must transform under the operations of such a group according to irreducible representations of this group (for points not along lines of symmetry, etc., the group may consist only of the identity operation). To each irreducible representation belongs an energy value. Following Bouckaert, Smoluchowski and Wigner's fundamental paper (1936) on the application of group theory to the classification of electronic levels in solids, it is customary to denote the particular point in $\mathbf{k}$-space by a capital Greek or italic letter (e.g. the point $k = 0$ is denoted by Γ), and the irreducible representation, and corresponding energy value and linear combination of basic functions is indicated by an arbitrary suffix; e.g. Γ_1 in cubic structures denotes that linear combination (of plane waves, of atomic orbitals, in general any basic functions) which is invariant under all the 48 operations of the cubic group. It seems more convenient to use as suffices the spectroscopic letters s, p, d ... to denote the first spherical harmonic which appears in a series development in terms of such functions. With this convention, one writes Γ_s instead of Γ_1. Some levels are degenerate, the number of states with the same energy being equal to the dimension of the corresponding irreducible respresentation. For instance, in cubic structures, at $k = 0$, Γ_p is triply degenerate, and there are two Γ_d levels, one doubly and one triply degenerate. In this example (Bethe 1929) the 'correspondence' with the atomic levels is very clear; it is easy to establish such a correspondence in every case, as well as a correspondence with the plane wave levels $E_n = |\mathbf{k}+\mathbf{K}_n|^2$. Apart from extending the group theoretical considerations to include spin–orbit interaction (Elliott 1954, Dresselhaus 1955), this is as far as one can go with purely symmetry considerations. Qualitatively it is quite far, since one can identify generally the most likely band structures (see, for instance, Glasser's (1959) discussion of the wurtzite structure). A comprehensive review of space groups and their representations has been given by Koster (1957). A general method for finding the wave functions belonging to a given irreducible representation has been given by Melvin (1956).

The next step in most methods is an application of the standard variational principle of quantum mechanics: one calculates the expectation value W of the

energy with (normalized) trial functions ψ_{trial}. Then

$$W = \int^{\infty} \psi^*_{\text{trial}} \mathscr{H} \psi_{\text{trial}} \, d\tau \qquad \ldots\ldots(2)$$

provides an upper limit for the lowest eigenvalue. As trial ψ one takes generally as many terms as one can deal with of an expansion in terms of functions χ_i with the correct symmetry for the level under examination, and satisfying the conditions of periodicity, chosen from a complete orthonormal set

$$\psi_{\text{trial}} = \sum_{i=1}^{n} a_i \chi_i. \qquad \ldots\ldots(3)$$

If the χ_i are atomic orbitals, the method goes under the name of tight binding ; if the χ_i are plane waves, one has ' momentum eigenfunctions ' methods. More often the χ_i are modified plane waves, e.g. ' augmented ' or ' orthogonalized '. Substituting (3) into (2), and using the notation

$$H_{ij} = \int^{\infty} \chi_i^* \mathscr{H} \chi_j \, d\tau,$$

one has

$$W = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i^* a_j H_{ij},$$

an expression which must be minimized with respect to the a_i's, subject however to the condition of normalization $\Sigma_i a_i^* a_i = 1$. Introduction of the Lagrangian multiplier λ leads to the system of equations $\Sigma_j a_j H_{ij} - \lambda a_i = 0$, $i = 1, 2, \ldots, n$ which are compatible only if

$$\begin{vmatrix} H_{11}-\lambda & H_{12} & H_{1n} \\ H_{21} & H_{22}-\lambda & H_{2n} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots \\ H_{n1} & H_{n2} & H_{nn}-\lambda \end{vmatrix} = 0. \qquad \ldots\ldots(4)$$

Equation (4) gives n real roots for λ and, as is well known, the lowest root gives the required upper limit to the lowest eigenvalue. With some methods the energy appears in (4) also implicitly, namely when the χ's have been obtained by numerical integration of Schrödinger's equation with a trial value of the energy. The higher roots are poor approximations to the higher energy levels of the same symmetry. If these are required, instead of increasing the number of terms in the expansion of ψ, it is much preferable to make first the trial function orthogonal to each of the wave functions associated with lower values of E. For instance, if the ψ's are plane waves, one would obtain the energy of the 1s electrons, unless one makes the trial function orthogonal to the ion core wave functions (orthogonalized plane waves).

In these methods one thus uses functions which are not solutions of Schrödinger's equation for the chosen potential, but which are continuous and satisfy automatically the periodicity condition

$$\psi(\mathbf{r} + \mathbf{R}_i) = \exp(i\mathbf{k} \cdot \mathbf{R}_i)\psi(\mathbf{r}). \qquad \ldots\ldots(5)$$

Another point of view is to take ψ's which are solutions of Schrödinger's equation, and devise methods for satisfying as accurately as possible the conditions of continuity and periodicity. This is the cellular method, in which the crystal is

divided into cells by drawing all planes bisecting and perpendicular to the lines joining each lattice site to its neighbours. In monatomic lattices all cells are identical. Each atom thus finds itself at the centre of a cell. In each of these the potential is assumed to be spherically symmetric (deviations from spherical symmetry may be considered later as a perturbation). Then the wave function may be separated in each cell in spherical polar coordinates, the angular part being expressed by means of associated spherical harmonics. Which spherical harmonics, or combinations of them, are present can be determined by group theory (see e.g. Von der Lage and Bethe 1947, Bell 1954). Each combination is multiplied by a radial function, determined by (generally numerical) integration of the radial Schrödinger equation with the given potential and trial values of the energy. For values of $\mathbf{k}$ along symmetry axes in $\mathbf{k}$-space the number of spherical harmonics which appear in the development of ψ is much reduced, thus it is convenient to carry out calculations for such points (this in fact is a general feature with all methods). One has then an expansion in each cell, and the main problem is to ' match ' these solutions, that is to satisfy the appropriate boundary conditions at the surfaces of the polyhedral cells. This of course is more complicated for polyatomic substances, and when the cell has a large number of faces and a low symmetry. The methods devised for this purpose will be examined in § 4.1. In particular, a variational procedure was developed to be used in the framework of the cellular approach (Kohn 1952, Jenkins and Pincherle 1954), which effects the best compromise between satisfying Schrödinger's equation and satisfying the boundary conditions, also eliminating arbitrariness in the choice of the approximate conditions to be used in the actual calculation. In practice, however, it has not proved to be of straightforward application. The assumption of spherical symmetry for the potential is not essential. Instead of separating out the angular part of the wave function, it is possible (Betts 1959) to construct polynomials in x, y, z having the correct symmetry and forming an orthonormal set in the polyhedron ; with these one can in principle deal with any form of potential.

An approach has also been developed which is not based on the differential form of Schrödinger's equation, but on the equivalent integral equation, the idea being that a solution by means of successive approximations may be more likely to succeed starting from the integral instead of the differential formulation (Dank and Callen 1952). Very akin to this method is an earlier one, that treats the problem as a scattering one (Korringa 1947).

Clearly then the fundamental methods for band structure calculation can be classified from different points of view, for instance, as already mentioned, into methods in which the correct periodicity is implicit from the start, and methods in which it is reached approximately at a later stage—cellular method (§ 4.1) ; or into methods using as basic functions some form of atomic orbitals, and methods using some form of plane waves.

Considering this last point of view, we have (§ 4.2.1), using atomic orbitals, the already mentioned tight binding approximation (the name is somewhat misleading : it should be reserved for the case when, using linear combination of atomic orbitals, only interaction with the nearest neighbours is taken into account), which was historically the first method suggested for deriving energy levels in solids (Bloch 1928). Among its modifications and improvements are (§ 4.2.2) the method of equivalent

orbitals (Hall 1952); (§ 4.2.3) the method of orthogonalized atomic orbitals (Löwdin 1950); (§ 4.2.4) methods based on the use of Wannier functions (Wannier 1937, Parzen 1953); and (§ 4.2.5) the semilocalized atomic orbitals (Morita 1958). In all these methods one then uses localized basic functions, for which the name 'site orbitals' has been suggested (Winston 1954).

With the other class of methods, the basic functions employed extend instead through the whole crystal. In the simplest case one uses (§ 4.3.1) just a linear combination of plane waves (momentum eigenfunctions). The most powerful modification of this method is (§ 4.3.2) the already mentioned method of orthogonalized plane waves (Herring 1940). There are also (§ 4.3.3) 'modified' plane waves (Brown and Krumhansl 1958). In another method based on plane waves (§ 4.4.1), the augmented plane waves (Slater 1953), one makes use of a further approximation, already considered by Slater in 1937, which has been widely applied: it assumes the potential to be constant outside atomic spheres, included or inscribed in the cellular polyhedron. Variations of the potential may be considered later when carrying out a variational calculation. The (§ 4.4.2) scattering matrix method (Korringa 1947) and the (§ 4.4.3) variational–iterational ones, based on the integral formulation of Schrödinger's equation, can be applied in practice only if use is made of this approximation (Kohn and Rostoker 1954). The approximation may be used also in conjunction with the cellular method (§ 4.4.4, Pincherle, unpublished).

Among the methods that make use of empirical parameters, one may quote (§ 4.5.1) the quantum defect method (Ham 1955, and references), which is actually not a method for working out band structures, but for expediting calculations; the already mentioned $\mathbf{k}.\mathbf{p}$ approximation (§ 4.5.2, Kane 1956), and (§ 4.5.3) Phillips' (1958) pseudo-potentials. To this class of methods belong also the attempts to derive, by a perturbation calculation, the band structure of the III–V intermetallic semiconductors, such as GaAs, from the known band structure of the Group IV elements, such as Ge (Herman 1955 b, Callaway 1957).

In the following section the main features of these various methods will be described, at least for the simplest cases, and the main calculations in which they have been applied will be mentioned. Some of the methods, and the main concepts involved in band structure calculations, are discussed in an instructive paper of Slater (1952) which works out the band structure of a fictitious crystal with a $\cos x + \cos y + \cos z$ potential.

§ 4. Details of the Various Methods

4.1. *The Cellular Method*

The main features of this method have already been described in § 3. It was first suggested by Wigner and Seitz (1933, 1934), and the polyhedra into which the crystal is divided are thus called Wigner–Seitz cells. To satisfy in a simple way the boundary conditions, the same authors suggested, for monatomic cubic crystals and $k = 0$, replacing the polyhedron with the sphere having the same volume. At $k = 0$ the s radial function, f_s, is the only important one, and the boundary conditions reduce to $(df_s/dr)_{r=R} = 0$, where R is the radius of the equivolume sphere. The condition will be satisfied only for certain values of the energy, and the number of nodes of the corresponding radial function decides

which of these values is the desired eigenvalue : for instance, if dealing with sodium, f_s must look like an atomic 3s function having its main maximum at $r = R$. The sphere approximation was useful in the early days, particularly for calculating cohesion energies. Bardeen (1938) suggested an interesting method for obtaining the eigenvalues of other states by perturbation theory ($\mathbf{k}.\mathbf{p}$ approach). One obtains $E(\mathbf{k})$ in the form $E_0 + E_2 k^2 + E_4 k^4 + \ldots$. A boundary condition method for obtaining the general coefficient E_{2n} was given by Silverman (1952). This approach is satisfactory for alkali metals (see, for instance, Callaway 1958 c).

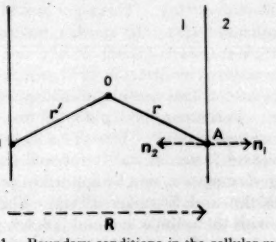


Figure 1. Boundary conditions in the cellular method.

In general however one wishes to satisfy the boundary conditions at the surfaces of the actual Wigner–Seitz polyhedra. Consider a monatomic substance : clearly it is then sufficient to consider just one polyhedron. At any point A (figure 1) on the surface one wishes ψ_1 in cell (1) to join smoothly to ψ_2 in cell (2), that is

$$\psi_1(\mathrm{A}) = \psi_2(\mathrm{A}) \; ; \quad \frac{\partial \psi_1}{\partial n_1}(\mathrm{A}) = -\frac{\partial \psi_2}{\partial n_2}(\mathrm{A}),$$

where ψ is developed in a series of spherical harmonics appropriate to the symmetry of the point $\mathbf{k}$ that one is considering, and the differentiation is along the outer normal to the corresponding cell. Since we wish to consider the wave function only in one cell, we introduce the point B on the opposite face of the polyhedron, which is brought into A by the fundamental translation $\mathbf{R}$. Then one must have, according to (5) :

$$\psi_1(\mathrm{A}) = \psi_1(\mathrm{B}) \exp(i\mathbf{k}.\mathbf{R}) \; ; \quad \frac{\partial \psi_1}{\partial n}(\mathrm{A}) = -\frac{\partial \psi_1}{\partial n}(\mathrm{B}) \exp(i\mathbf{k}.\mathbf{R})$$

and two equations of this type should be satisfied at all points on the surface, requiring an infinite number of coefficients in the development of ψ. Clearly one will have to resort to approximations. In any case, the system of equations for such coefficients will have non-trivial solutions only for certain values of the energy which are the required eigenvalues.

Using then a finite number of terms in the expansion of ψ, the first approximate method that comes to mind is to satisfy the boundary conditions at a sufficient number of chosen points. This procedure has been used widely, the chosen points being generally the centres of the faces, or points on the intercept of the equivolume sphere, but unfortunately the energy values turn out to be rather sensitive to the

choice of the matching points, a fact that makes the results of many of the early calculations open to doubt. The point matching procedure was systematized by Howarth and Jones (1952), who suggested taking sets of points and increasing the number of terms in the expansion of ψ until the energy value is the same, within the required approximation, for all sets. A similar criterion is now used in general for testing the convergence of the results. Altmann (1958) has used a least squares procedure that has proved satisfactory : if the number of coefficients c_i one wishes to consider is m, the equations are of the form

$$\sum_{i=1}^{m} a_{ri} c_i = 0 ;$$

choose then a large number n of points and minimize the sum of the squares of all the residuals

$$S^2 = \sum_{r=1}^{n} \left[\sum_{i=1}^{m} a_{ri} c_i \right]^2 .$$

All the angular and radial functions in the expansion of ψ are taken to be normalized, and S^2 is then to be minimized subject to the condition

$$\sum_{i=1}^{m} c_i^2 = 1.$$

One introduces a Lagrangian multiplier in the usual way, and the procedure is then formally identical to that for the variational problem discussed in § 3.

It has also been suggested (Bell *et al.* 1953), particularly for polyatomic lattices, to treat the problem as the matching of distributions on the surface of the polyhedron ; that is, of the distribution of ψ and $\partial\psi/\partial n$ calculated in the two adjacent cells. One can then for instance equate average values and moments of these distributions.

There are still other methods (cf. Pincherle 1958), the most satisfactory being in principle the already quoted variational procedure suggested by Kohn (1952). One considers, for monatomic lattices, the functional

$$J = \mathscr{R}\left[\int_{\substack{\text{volume of}\\ \text{unit cell}}} \psi^*(\mathscr{H} - E)\psi \, d\tau + \int_{\substack{\text{surface of}\\ \text{unit cell}}} \psi^*(\mathbf{r}') \exp\left[i\mathbf{k}.(\mathbf{r}' - \mathbf{r})\right] \frac{\partial\psi(\mathbf{r})}{\partial n} \, d\sigma \right] \qquad \ldots\ldots(5')$$

where $\mathbf{r}$ is any point on the surface of the Wigner–Seitz cell and $\mathbf{r}'$ is a conjugate point such that $\mathbf{r}' - \mathbf{r}$ is a fundamental translation $\mathbf{R}$ (figure 1), and $\mathscr{R}$ means ' real part of '.

J vanishes for the correct $\mathbf{k}$, E, ψ, thus the method may be used, if one has approximate ψ's, to find the relation between E and $\mathbf{k}$ from $J = 0$; the error in E is of the order of the square of the error in ψ. But, more important, J has the property of being stationary with respect to arbitrary variations of ψ when ψ is the solution of the boundary value problem. Therefore a variational procedure is most powerful. Take as usual $\psi = \Sigma_{i=1}^{n} c_i \chi_i$, where the χ_i are suitably chosen to fit the case under examination as explained in § 3 ; substitute into (5′) and demand that $\partial J/\partial c_i = 0$ for all c_i. A system of linear equations is obtained for the c_i's : the vanishing of the determinant of the coefficients gives the relation between E and $\mathbf{k}$.

Contrary to what happens in the ordinary Rayleigh–Ritz procedure, the energy values so determined do not represent upper bounds to the true eigenvalues.

How (5′) should be best applied, also in the case of polyatomic lattices, has been discussed by Jenkins and Pincherle (1954).

An interesting idea that many have tried, unsuccessfully so far, to develop is that first put forward by Wannier (1938) of introducing ' coordinates to fit the cell ', for instance in simple cubic crystals a system of coordinates in which the surfaces ' r ' = constant are spheres near the origin, but approach cubes farther out. Such a system of coordinates would practically remove the difficulties in satisfying the boundary conditions.

Among the most extensive calculations carried out with the cellular method are that of Howarth and Jones (1952) on Na, that of Bell, Hum, Pincherle, Sciama and Woodward (1953) on PbS (in which an attempt was made to establish a self-consistent field by making self-consistent the charge distribution for $k = 0$), that of Howarth (1953) on Cu, those of Schiff on Li (1954) and on Ti (1955), that of Jenkins (1954, 1956) on Si, that of Altmann (1958) on Zr, that of Altmann and Cohan (1958) on Ti, and that of Shakin and Birman (1958) on ZnS. Birman (1959) has also worked out with the cellular method the order of the levels in the two modifications (wurtzite and zincblende) of ZnS.

The cellular method was until a few years ago the most widely used. At the moment it has fallen into disfavour, mainly because : (*a*) it is restricted in practice to spherically symmetric potentials ; (*b*) there are still difficulties connected with satisfying the boundary conditions ; (*c*) one must keep a large number of terms in the expansion of ψ for really accurate results ; (*d*) it is somewhat less suited to automatic computers than other methods. It is, however, one of the most natural methods, and one in which the wave functions have a familiar form.

4.2. *Methods based on Atomic Orbitals*

4.2.1. *Tight binding.*

In the paper (1928) in which he established for the first time the form of the wave functions in crystals, Bloch suggested also that one may approximate to them by a linear combination of atomic orbitals (LCAO) of the form

$$\psi_{\mathbf{k}}(\mathbf{r}) = N \sum_{\mathbf{R}_i} \exp(i\mathbf{k} \cdot \mathbf{R}_i)\, \phi(\mathbf{r} - \mathbf{R}_i), \qquad \ldots\ldots(6)$$

where N is a normalizing factor and $\phi(\mathbf{r} - \mathbf{R}_i)$ is an atomic wave function (for instance, if dealing with sodium it may be a 3s function) centred at a lattice site $\mathbf{R}_i$. Hybrid orbitals may also be used (Flodmark 1959). It is easy to see that (6) satisfies the fundamental periodicity condition (5). The method will be satisfactory if the crystal wave function behaves in fact much like an atomic function near each atom ; this is clearly the case if the electron is tightly bound, hence the name ' tight binding ' for the method. One problem for which the method may be expected to be appropriate is that of the d-bands in transition metals.

Indicating by $H(\mathbf{R}_n)$ the matrix element of the Hamiltonian $\mathscr{H}$ between two atomic orbitals located on atoms separated by $\mathbf{R}_n$,

$$H(\mathbf{R}_n) = \int^{\infty} \phi^*(\mathbf{r})\, \mathscr{H}\, \phi(\mathbf{r} + \mathbf{R}_n)\, d\tau$$

and by $S(\mathbf{R}_n)$ the ' overlap integral ' between such orbitals

$$S(\mathbf{R}_n) = \int^{\infty} \phi^*(\mathbf{r})\,\phi(\mathbf{r}+\mathbf{R}_n)\,d\tau,$$

the average value of the energy $\bar{E}$ with wave functions (6) (upper limit to the eigenvalue) is given by

$$\bar{E} = \frac{\sum_{\mathbf{R}_n} \exp(-i\mathbf{k}.\mathbf{R}_n) H(\mathbf{R}_n)}{\sum_{\mathbf{R}_n} \exp(-i\mathbf{k}.\mathbf{R}_n) S(\mathbf{R}_n)}. \qquad \ldots\ldots(7)$$

The integrals occurring in (7) are of a type well known in molecular theory. $H(\mathbf{R}_n)$ is a ' three centres integral ' : $\phi^*(\mathbf{r})$ is centred about one atom, $\phi(\mathbf{r}+\mathbf{R}_n)$ about another, and H involves potentials centred about third atoms, since the crystal potential is generally expressed as a superposition of atomic potentials. Although with the use of automatic computers much progress has been made in the evaluation of such integrals, their calculation is still laborious. And of course, the use of (7) is only a first step. For an accurate calculation one must consider wave functions made up of several ψ's of the type of (6) corresponding to different atomic orbitals (e.g. s, p, d...) and use them in a variational procedure : the trial wave function has then the form of a Slater determinant of atomic orbitals. Thus for an accurate determination of band structures the method appears inferior to others, even with the refinements that have been suggested. On the other hand, simplified treatments, such as using simply equation (7) (for non-degenerate levels—and as few orbitals as possible for degenerate bands) and neglecting all interactions except those between nearest neighbours, are easy to carry out, are instructive and give a qualitative idea of the band structure. The magnitude of the overlap integrals is a criterion for the applicability of the method : they must be small, no larger than, say, 0·1, as compared with one for maximum overlap. As is well known, small overlap means narrow bands.

Slater and Koster (1954) have also suggested using the LCAO method as an interpolation procedure, when some energy levels are known, either empirically or through a computation carried out by other methods. The parameters, essentially the $H(\mathbf{R}_n)$, appearing in the LCAO formulae are found from these known levels and then the formulae may be used to give approximate values of the energy at all other points in **k**-space. This is useful because, as has already been mentioned, one generally determines the value of E at particular points of symmetry. As we shall see, other interpolation procedures have been suggested, but this is probably the easiest. Whether it gives better results than any reasonable empirical interpolation formula has not been proved, and in fact in the case of silicon (Bassani 1959 b) it has been shown not to be very reliable.

The tight binding method, by its nature, is not suited for excited states (empty levels). It has been shown by Parmenter (1952) that it can be modified to converge more rapidly, but it then becomes identical with the method of orthogonalized plane waves.

As with the cellular method, many of the early results obtained with the tight binding procedure suffer from over-simplification. Among recent applications are Casella's calculation on NaCl (1956), and Bilz's (1958) on NaCl structures, typical of the use of a simplified version of the method to obtain indications on a band structure ; Fletcher and Wohlfarth's (1951) (Fletcher 1952) calculation on the

d-bands of Ni ; Grimley's (1958) study of the valence band of LiF in which, to allow for a partly homopolar binding, the LCAO consisted of the outer orbitals on the negative ion and the first vacant orbitals on the positive one ; Mase's (1958) calculation on Bi ; Howland's (1958) treatment of the valence band of KCl, in which to avoid the possibility of the band structure depending on the details of the one-electron potential, all the interactions were calculated directly from many-centre integrals between the atomic orbitals. This calculation was performed fairly accurately for the experimental atomic distance a, and also more approximately for other values, to determine how the various levels shift when a changes. This last is another type of problem for which the tight binding procedure gives useful indications with a fairly simple calculation. Other applications of tight binding include García Moliner's (1958 b) demonstration that Pippard's (1957) experimental results on the shape of the Fermi surface in copper are fitted to 1% accuracy by an expression resulting from a tight binding calculation with second neighbour interactions ; Reitz's (1957) study of the band structure of Se and Te ; Slonczewski and Weiss' (1958) calculation on graphite in which, as in earlier treatments, a two-dimensional model was used, the interaction between layers being considered later as a perturbation.

4.2.2. *Equivalent orbitals.*

This modification of the Bloch method is based on the observation that, if one has a crystal wave function represented by a Slater determinant (completely antisymmetric combination) of atomic orbitals, the values of the energy, distribution of charge density, etc., are not changed if one applies a unitary transformation to the elements of the determinant. In crystals such as diamond, where the probability of finding the valence electrons peaks sharply in the regions of the tetrahedral bonds, there may be an advantage in using orbitals which are localized at these bonds. These are the equivalent orbitals, obtained from the atomic orbitals through a unitary transformation. Mathematically, the equivalent orbitals, $\beta_i(x_j)$ are in principle solutions of the self-consistent field equations and may be defined by the condition that their self-energy

$$\sum_i \frac{\beta_i^*(1)\beta_i(1)\beta_i^*(2)\beta_i(2)}{r_{12}}$$

is a maximum with respect to unitary transformations of occupied orbitals. Since the self-energy is a measure of localization, the equivalent orbitals are the most localized orbitals. These functions have found extensive application in the theory of molecules (see for instance Hall 1959, and references), but in crystals their application has been explored only for diamond, silicon, germanium (Hall 1958) and for body-centred cubic metals (Hall 1953).

Also in this method, one may introduce interactions of distant bonds, or only interactions between nearest bonds. Further, the parameters appearing in the formulae may be obtained empirically and then the formulae used for interpolation. The extension of the concept to unoccupied states has met with difficulties. The advantages that the equivalent orbitals present over the atomic orbitals have been discussed by Hall (1958). One of these advantages is the avoidance of the difficulties connected with the orthogonalization of the orbitals centred on different atoms.

4.2.3. *Orthogonalized atomic orbitals* (Löwdin 1950).

The atomic orbitals centred on different atoms are not orthogonal and therefore one needs to consider the overlap integrals. One can however take linear combinations such that the functions centred on different atoms are orthogonal. If S_{ij} is the overlap matrix between the atomic orbitals ϕ,

$$S_{ij} = \int^{\infty} \phi^*(\mathbf{r} - \mathbf{R}_i)\,\phi(\mathbf{r} - \mathbf{R}_j)\,d\tau - \delta_{ij},$$

then Löwdin showed that the

$$\chi(\mathbf{r} - \mathbf{R}_i) = \sum_j \phi(\mathbf{r} - \mathbf{R}_j)\,(1 + S)_{ji}^{-1/2}$$

are the required orthogonalized atomic orbitals. Expanding the square root :

$$\chi(\mathbf{r} - \mathbf{R}_i) = \phi(\mathbf{r} - \mathbf{R}_i) - \tfrac{1}{2}\sum_j \phi(\mathbf{r} - \mathbf{R}_j)\,S_{ji} + \tfrac{3}{8}\sum_j \sum_m \phi(\mathbf{r} - \mathbf{R}_j)\,S_{jm}\,S_{mi} + \ldots .$$

The crystal wave functions may now be constructed from the χ's in the same way as from the ϕ's. As in equation (6) the coefficients are $\exp(i\mathbf{k}.\mathbf{R}_i)$. Note that, as with the ϕ's, the crystal functions

$$\psi_{\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{R}_i} \exp(i\mathbf{k}.\mathbf{R}_i)\,\chi(\mathbf{r} - \mathbf{R}_i)$$

are only approximate. The variational procedure is then formally similar to that in the original tight binding method. Löwdin (1951, 1956) succeeded also in reducing all the integrals occurring to two-centre integrals.

Although this method has the advantage of eliminating overlap integrals, working with orthogonalized atomic orbitals is complicated. Slater (1951 a) has argued that they are not superior to the LCAO's while Raimes (1954 b) has stressed that the error introduced by the fact that in practice Löwdin's functions are not solutions of the self-consistent field equations for the valence electrons in the crystal may be serious. Thus, despite its mathematical elegance, this method, primarily designed for the evaluation of cohesive energies, does not seem very suitable for band structure calculations, except for tightly bound electrons. It was applied by Ariyama and Mase (1954) to a study of the electronic structure of graphite and boron nitride.

4.2.4. *Methods based on the use of Wannier functions.*

Considering Bloch's formula

$$\psi_{\mathbf{k}}(\mathbf{r}) = N\sum_{\mathbf{R}_i} \exp(i\mathbf{k}.\mathbf{R}_i)\,\phi(\mathbf{r} - \mathbf{R}_i) \qquad \ldots\ldots(6)$$

we have already remarked that, neither when the ϕ's are atomic orbitals, nor when they are equivalent orbitals, nor when they are orthogonalized atomic orbitals, is ψ the exact solution of Schrödinger's equation for the crystal. The functions $W(\mathbf{r} - \mathbf{R}_i)$ which, introduced into (6) instead of the ϕ's, do make ψ the exact solution of Schrödinger's equation are the Wannier functions (Wannier 1937). Since (6) is a Fourier expansion, one can conversely express the Wannier functions in terms of the ψ's :

$$W(\mathbf{r} - \mathbf{R}_i) = \frac{1}{\Omega_B}\int \psi(\mathbf{r}, \mathbf{k}) \exp(-i\mathbf{R}_i.\mathbf{k})\,d\mathbf{k} \qquad \ldots\ldots(8)$$

the integration being extended to the unit cell of k-space (first Brillouin zone) of volume Ω_B, namely over all the states of one band. One Wannier function thus represents a complete band. At large interatomic distances the Wannier functions coincide with the atomic orbitals. If the ψ's are normalized to one in the unit cell, the Wannier functions defined by (8) are normalized to unity over the whole space.

The Wannier functions satisfy the system of differential equations (Koster 1953)

$$\mathscr{H} W(\mathbf{r}-\mathbf{R}_j) = \sum_{\mathbf{R}_i} W(\mathbf{r}-\mathbf{R}_j-\mathbf{R}_i)\,\epsilon(\mathbf{R}_i) \qquad \ldots\ldots(9)$$

in which only the Wannier functions of the band under consideration appear, and

$$\epsilon(\mathbf{R}_i) = \frac{1}{\Omega_B}\int_{\substack{\text{unit cell of}\\ k\text{-space}}} E(\mathbf{k})\exp(-i\mathbf{R}_i.\mathbf{k})\,d\mathbf{k} \qquad \ldots\ldots(10)$$

are the Fourier coefficients of the energy $E(\mathbf{k})$ (which is a periodic function in momentum space). These Fourier coefficients of the energy are also the matrix elements of the Hamiltonian between Wannier functions centred at different lattice sites:

$$\epsilon(\mathbf{R}_j-\mathbf{R}_i) = \int^{\infty} W^*(\mathbf{r}-\mathbf{R}_i)\,\mathscr{H} W(\mathbf{r}-\mathbf{R}_j)\,d\tau. \qquad \ldots\ldots(11)$$

The Wannier functions centred on different lattice sites are all identical, thus it is only necessary to calculate $W(\mathbf{r})$. They are also orthogonal, thus avoiding the difficulties that in the Bloch approximation arise through lack of orthogonality. It has also been shown (Carlson and Keller 1957) that the Wannier functions W_i approximate the atomic orbitals ϕ_i more closely in a least squares sense

$$\left(\text{i.e.}\quad \sum_i \int^{\infty} |W_i-\phi_i|^2\,d\tau \quad \text{is a minimum}\right)$$

than do the functions of any other orthonormal set.

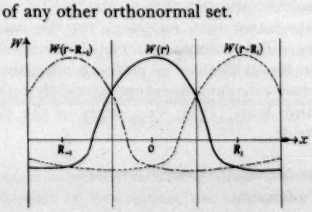


Figure 2. Wannier functions (schematic).

One can think of the Wannier functions either as a single function $W(\mathbf{r})$ (one for each band) extending to the whole of space, or as a set of functions $W(\mathbf{r}-\mathbf{R}_i)$ defined only in one cell and being the Fourier coefficients of the expansion of ψ in plane wave in k-space. Figure 2 gives a schematic representation of the Wannier functions along an arbitrary direction x. $W(\mathbf{r}-\mathbf{R}_i)$ is displaced by $\mathbf{R}_i$ with respect to $W(\mathbf{r})$.

Because of these interesting properties the Wannier functions are very useful for discussing the electronic properties of solids, and in principle it is convenient to have a single function completely describing a whole band. If this function is

known, one gets ψ for any $\mathbf{k}$ from equation (6), and the Fourier coefficients of the energy, $\epsilon(\mathbf{R}_i)$, from equation (11). Then the $E(\mathbf{k})$ relation is given by

$$E(\mathbf{k}) = \sum_{\mathbf{R}_i} \epsilon(\mathbf{R}_i) \exp(i\mathbf{R}_i.\mathbf{k}). \qquad \ldots\ldots(12)$$

However the system of differential equations (9) is not easy to solve and no direct determination of the Wannier functions from these equations has been made so far. Of course the equations (9) cannot alone give all the $\epsilon(\mathbf{R}_i)$: one must add other conditions, and it is clear that these must be the conditions of orthogonality between functions centred on different lattice sites. Parzen (1953) and Koster (1953) have succeeded in formulating the problem as a variational principle, and Wainwright and Parzen (1953) have applied the procedure to lithium. Results were in agreement with those of other methods. However, in constructing the trial functions it was found more advantageous to use the Fourier transform of the Wannier function, which (§ 4.3.1) is the momentum eigenfunction : in effect one had then a plane wave procedure. A variational calculation with Wannier functions has been carried out also for sodium (Miasek 1956).

4.2.5. *Semi-localized crystal orbitals* (Morita and Takahashi 1958).

Originally this method was developed to take into account the effect of homopolar binding in determining the cohesion energy of ionic crystals. It started from the point of view of the valence bond, rather than of the molecular orbitals model, but its formalism is very similar to that of the other methods examined in this Subsection. An example will make clear what these orbitals are : consider the LiH crystal : with a tight binding procedure one would place 1s orbitals on the H^- ion. Here one takes a linear combination of such a 1s orbital and of 2s orbitals on the six Li^+ ions nearest to the H^- ion considered. In this way one can deal with resonance between two structures using a single Slater determinant.

The method is primarily one for the calculation of cohesive energies, and indeed the effective one-electron Hamiltonian and its matrix elements, required for band structure calculations, are derived from the computation of the cohesive energy. This has the advantage of eliminating ambiguities due to the choice of the potential (there have been other attempts along these lines, e.g. Howland's (1958)). If the cohesive energy has not been calculated, the semi-localized crystal orbitals formulae may serve as an interpolation procedure, the parameters appearing in them being determined by comparison with experiment. This has been done by Morita (1958, 1959) for Si and Ge.

In this method there are no difficulties over orthogonality ; and the conduction and valence bands are determined by independent secular equations, which are thus of a smaller order. On the other hand, the method cannot be applied to metals.

4.3. *Plane Wave Methods*

4.3.1. *Simple plane waves.*

Since, as is well known, $\psi_{\mathbf{k}}(\mathbf{r})$ has the form of the product of a periodic function, with the periodicity of the lattice, times the factor $\exp(i\mathbf{k}.\mathbf{r})$, it can be expanded in a Fourier series (in a series of plane waves),

$$\psi_{\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{K}_n} v(\mathbf{k}+\mathbf{K}_n) \exp[i(\mathbf{k}+\mathbf{K}_n).\mathbf{r}] \qquad \ldots\ldots(13)$$

where the sum is over all fundamental vectors of the reciprocal lattice. The coefficients

$$v(\mathbf{k}+\mathbf{K}_n) = \frac{1}{\Omega}\int_{\text{unit cell}} \psi_{\mathbf{k}}(\mathbf{r}) \exp\left[-i(\mathbf{k}+\mathbf{K}_n).\mathbf{r}\right] d\mathbf{r} \qquad \ldots\ldots(14)$$

where Ω is the volume of the unit cell of the direct lattice, are called momentum eigenfunctions, since they are clearly the probability amplitudes for the various possible values of momentum $\mathbf{k}$, $\mathbf{k}+\mathbf{K}_n$, etc. Only a discrete set of values of momentum is possible for an electron in a given state $\mathbf{k}$. In a state with momentum $\mathbf{k}+\mathbf{K}_n$, a free electron has energy $|\mathbf{k}+\mathbf{K}_n|^2$. If ψ does not deviate much from a plane wave (nearly free electrons), only $v(\mathbf{k})$ is large. If we assume that all the v's are negligible, except one, we have the model of free electrons.

If we let $\mathbf{k}$ take all values inside the first Brillouin zone, all the coefficients vary and are continuous functions of $\mathbf{k}$. However, as with the Wannier functions in ordinary space (figure 2), they all are in fact the same function of $\mathbf{k}$, only displaced by fundamental vectors of the reciprocal lattice. We can thus speak of one function $v(\mathbf{k})$, localized on the reciprocal lattice site $\mathbf{k}=0$. Of course, there are different v functions for different bands. If the band is narrow and deviates little from an atomic level, v approaches the probability amplitude of momentum for the corresponding atomic state; if we have free electrons, only one of the coefficients is different from zero in a zone, and is constant through the zone. For a discussion of some aspects of the momentum distribution in solids, see Donovan and March (1956). For any band, the Wannier function $W(\mathbf{r})$ and the momentum eigenfunction $v(\mathbf{k})$ are Fourier transforms one of the other:

$$v(\mathbf{k}) = \frac{1}{\Omega}\int^{\infty} W(\mathbf{r}) \exp\left(-i\mathbf{k}.\mathbf{r}\right) d\mathbf{r}. \qquad \ldots\ldots(15)$$

If the ψ's are normalized to one within the unit cell of volume Ω, then

$$\int^{\infty} v^*(\mathbf{k})\, v(\mathbf{k})\, d\mathbf{k} = \frac{8\pi^3}{\Omega^2}.$$

To find the system of equations satisfied by the v's, one must expand also the potential V (which is always taken to have the same periodicity of the lattice) in a Fourier series

$$V(\mathbf{r}) = \sum_{\mathbf{K}_n} \phi(\mathbf{K}_n) \exp\left(i\mathbf{K}_n.\mathbf{r}\right). \qquad \ldots\ldots(16)$$

Substituting then (13) and (16) into Schrödinger's equation $-\nabla^2\psi + V\psi = E\psi$, multiplying by $\exp\left[-i(\mathbf{k}+\mathbf{K}_i).\mathbf{r}\right]$ and integrating over unit cell, one obtains, since the plane waves $\exp\left(i\mathbf{K}_j.\mathbf{r}\right)$ are orthogonal in the unit cell,

$$\left[|\mathbf{k}+\mathbf{K}_i|^2 - E\right] v(\mathbf{k}+\mathbf{K}_i) + \sum_{\mathbf{K}_n} \phi(\mathbf{K}_n)\, v(\mathbf{k}+\mathbf{K}_i-\mathbf{K}_n) = 0, \qquad \ldots\ldots(17)$$

a system of purely algebraic equations for the v's, in which the Fourier coefficients of the potential, $\phi(\mathbf{K}_n)$ are the known quantities. The standard variational procedure leads to the same equations. The eigenvalues E can be found by successive approximations, solving the secular equation for E with more and more terms included in the summation. The convergence depends on how many of the ϕ's have an appreciable value.

This procedure is the straightforward extension of the usual procedure for solving the Mathieu equation, which corresponds to a one-dimensional lattice with a $\cos x$ potential. Discussing this simple example will make the method clearer. The period of this lattice is 2π, that of the reciprocal lattice is 1, thus $\mathbf{K}_n = n$. The Fourier coefficients of the potential are $\phi_1 = \phi_{-1} = \frac{1}{2}$, all others are equal to zero. Considering for instance $k = 0$, the system (17) reduces to

$$(n^2 - E)\,v(n) + \tfrac{1}{2}v(n+1) + \tfrac{1}{2}v(n-1) = 0.$$

The determinant of the coefficients of the v's must vanish, and thus one has the secular equation :

$$\begin{vmatrix} \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & 4-E & \frac{1}{2} & 0 & 0 & 0 & \dots \\ \dots & \frac{1}{2} & 1-E & \frac{1}{2} & 0 & 0 & \dots \\ \dots & 0 & \frac{1}{2} & -E & \frac{1}{2} & 0 & \dots \\ \dots & 0 & 0 & \frac{1}{2} & 1-E & \frac{1}{2} & \dots \\ \dots & 0 & 0 & 0 & \frac{1}{2} & 4-E & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \end{vmatrix} = 0. \quad \dots\dots(18)$$

The 5×5 determinant written explicitly in (18) already gives the lowest eigenvalue correctly to four figures and the next two eigenvalues to three figures, thus the procedure in this case converges rapidly not only for the lowest eigenvalue, but also for the first excited states. In three dimensions, the rows and columns of the determinant must correspond not to single plane waves, but to combinations of plane waves of the correct symmetry.

However one may anticipate that, since electrons in solids are in most cases represented by wave functions which oscillate rapidly near the nuclei, thus introducing high Fourier components, very many plane waves will be necessary, and the secular equation will be of very high order. Moreover, one would obtain most accurately the lowest eigenvalues, corresponding to inner core electrons, which are not of interest. We shall examine in the following Sections the modifications of the method that have made it suitable for the energy states of interest in solids. Nevertheless, combinations of simple plane waves with the correct symmetry have been used with fairly satisfactory results in special problems, e.g. by Ten Seldam (1957) and by Behringer (1959) to determine for which lattice distance solid helium and LiH, respectively, would become metals. They have been used also by Callaway (1959) for a study of the d-bands of body-centred cubic metals, employing a model of these crystals.

The extreme approximation of representing the electrons by one single plane wave is useful to provide a first picture of the bands of a crystal with a certain structure. This is the free electron model, sometimes used also to check the validity of approximations in band calculations, and then called ' empty lattice test ' (Shockley 1937). The energies are $E_n = |\mathbf{k} + \mathbf{K}_n|^2$ and the bands can thus be drawn without any further calculation. For instance, suppose that the first Brillouin zone is the truncated octahedron illustrated in figure 3, as is the case for metals with a face-centred cubic structure, e.g. Al, for substances with the diamond structure, like Si and Ge, for zinc-blende structures, for instance most of the III–V intermetallic semiconductors, and for crystals with the NaCl structure.

The bands according to the free electron model, along the four crystallographic directions (0, 0, 1) (Γ to X); (1, 1, 0) (Γ to K); (1, 1, 1) (Γ to L); (0, 2, 1) (Γ to W) are illustrated in figure 4. Numbers on the diagram indicate the degeneracy of the corresponding levels. The lowest branch is part of the free electron 'parabola' $E = p^2/2m$ centred on the cell under consideration; the other branches are parts of 'parabolae' centred on equivalent reciprocal lattice sites.

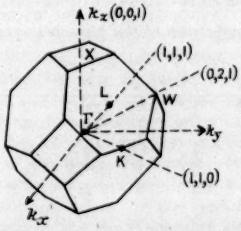


Figure 3. First Brillouin zone for face-centred cubic structures.

The position of the Fermi level is also indicated as a function of the number of electrons per unit cell, N. For example, for Al $N = 3$, and one sees immediately that the first band must be completely full, except perhaps at W, and that there may be electrons in the third band in the region of **k**-space near K and W. For

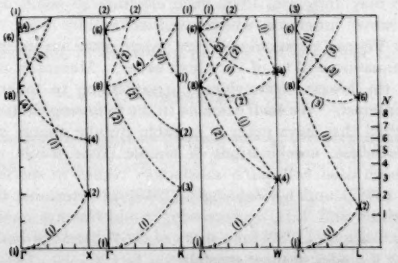


Figure 4. Energy bands in the free electron model for face-centred cubic structures, along four crystallographic directions.

Si and Ge $N = 8$, since there are two atoms per unit cell. The shape of the Fermi surface in metals has also been found to approach closely that given by this simple model (see e.g. Gold 1958).

The usefulness of these simple pictures as a basis for discussing any band structure has been demonstrated for instance by Herman (1958). The model may be considerably improved in a simple way by taking into account correlation interactions between the carriers with the method of Bohm and Pines (Pines 1955, Raimes 1954 a, Heine 1957).

By the methods of group theory it is straightforward to find in which crystal levels each free electron level splits when the rotational symmetry of the lattice and the value of **k** are taken into consideration (free electron correspondence). For instance, the 8-times degenerate level at $k = 0$ shown in figure 4 splits into four levels, Γ_s, Γ_f (non-degenerate) and Γ_p, Γ_d (each triply degenerate). One could also in principle say something about the relative position of these levels, if one had some idea of the Fourier coefficients of the potential. This further step is referred to as the approximation of ' nearly free electrons ', which proved very fruitful in the early study of metals and alloys. With this perturbation method, if one has two or more plane waves which have the same energy at some point inside or at the boundary of the first Brillouin zone, one sets up those linear combinations of the waves that make the energy diagonal, ignoring all other waves with different energies ; that is, one solves (18) considering only those v's whose arguments are equivalent points of **k**-space. The roots of the secular equation will in general be different, though the degeneracy may not be completely removed. For instance, considering, for a face-centred cubic structure, the eight waves degenerate at the second level at $k = 0$, one finds the roots

$$\begin{aligned}
E_1 &= E_0+\phi_1-\phi_2-\phi_3 \quad \text{(three times degenerate)}\\
E_2 &= E_0-\phi_1-\phi_2+\phi_3 \quad \text{(three times degenerate)}\\
E_3 &= E_0+3\phi_1+3\phi_2+\phi_3 \quad \text{(non-degenerate)}\\
E_4 &= E_0-3\phi_1+3\phi_2-\phi_3 \quad \text{(non-degenerate)},
\end{aligned}$$

where E_0 is the unperturbed energy, ϕ_1 is the Fourier coefficient of the potential associated with reciprocal lattice vectors of the type $(2, 0, 0)$, ϕ_2 that associated with vectors of the type $(2, 2, 0)$, and ϕ_3 that associated with vectors of the type $(2, 2, 2)$.

Substituting back into the equations (17), one gets for each value of the energy the corresponding combination of plane waves. On the other hand, from group theory one can find the linear combinations of plane waves corresponding to the levels Γ_s, Γ_p, Γ_d, Γ_f. By comparison one can then identify E_1 with Γ_p, E_2 with Γ_d, E_3 with Γ_s and E_4 with Γ_f. Thus, if $|\phi_1| > |\phi_2| > |\phi_3|$, and all coefficients are negative, the order of the levels may be guessed as Γ_s (lowest), Γ_p, Γ_d, Γ_f. This however assumes that the average potential is the same for all states, which is not generally the case. As with other simple procedures, also this method can in practice be applied only in connection with experimental data, or other theoretical results. It has been applied also to the density of states in the first few bands of body-centred cubic crystals (Nutkins 1956).

4.3.2. *Orthogonalized plane waves* (Herring 1940).

A very detailed discussion of this method has been given by Woodruff (1957). In brief, one takes as basic functions χ_i, to be used in the variational procedure (3), plane waves which have been made orthogonal to the core wave functions, for which one assumes that the tight binding approximation is satisfactory. If the atomic orbitals are u_{nlm}, one writes then for the core wave function of wave vector **k** :

$$\psi_{nlm,\,\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{R}_i} u_{nlm}(\mathbf{r}-\mathbf{R}_i)\exp(i\mathbf{k}\,.\,\mathbf{R}_i). \qquad \ldots\ldots(19)$$

For the valence electrons, or the unoccupied states, one takes

$$\chi = \exp(i\mathbf{k}.\mathbf{r}) - \sum_{n,l,m} B_{nlm,\mathbf{k}}\,\psi_{nlm,\mathbf{k}} \qquad \ldots\ldots(20)$$

the coefficients B being determined so as to make χ orthogonal to all the occupied core states ψ_{nlm}. One finds that they are in fact independent of the magnetic quantum number m, and given by

$$B_{nl,\mathbf{k}} = [4\pi(2l+1)]^{1/2}\, i^l \int_0^\infty r R_{nl}(r)\, j_l(kr)\, dr$$

(apart from a constant depending on normalization), where $R_{nl}(r)$ is the radial part of u, and $j_l(kr)$ is the spherical Bessel function of order l. Out of the χ's (20) one forms ψ_{trial}, taking only equivalent waves $\mathbf{k}+\mathbf{K}_n$, $\mathbf{K}_n$ being a fundamental vector of the reciprocal lattice. For each value of $\mathbf{k}$ one will take the linear combination of plane waves appropriate to the state under study, e.g. $\Gamma_s, \Gamma_p \ldots$, as discussed in § 4.3.1. One solves the secular equation adding more and more orthogonalized plane waves until adding new waves does not change the value of E to the required approximation. The different eigenvalues of the secular equations correspond to successive bands. Note that in this method it is not necessary to make special assumptions on the potential, for instance that it is spherically symmetric, since the potential is expressed through the Fourier coefficients (although spherical symmetry is assumed when calculating the coefficients $B_{nl,\mathbf{k}}$). A method for calculating these Fourier coefficients when the valence electron wave functions are expressed as combinations of orthogonalized plane waves has been given by Callaway and Glasser (1958).

Besides this advantage, the orthogonalized plane wave method is probably best suited for attempting to establish a self-consistent field, and for examining the effect of the potential on the band structure. For instance, Callaway (1958 b) used for Na a potential $V = 2/r + (A/r)\exp(-br)$, where the parameters A, b were taken from spectroscopic data: it is clear that a systematic study is here possible of the changes in band structure produced by changing the values of A and b. The orthogonalized plane wave procedure is also convenient for programming for an automatic computer.

Among its difficulties, one is that the usually available atomic orbitals assume a different potential for each state, and thus are not orthogonal. To avoid complications, it is then better to use atomic functions which are eigenfunctions of the same Hamiltonian used for the valence electrons. Another difficulty is that at $\mathbf{k}$-points of high symmetry the wave functions belonging to some of the irreducible representations of the group of $\mathbf{k}$ may already be orthogonal to all the Bloch atomic orbitals of the same $\mathbf{k}$. Since a simple plane wave expression is so slowly convergent, Callaway (1955 a) has suggested adding in this case a single cellular function of the correct symmetry and vanishing at the boundary of the atomic cell, to reproduce the correct behaviour of the wave function near the nucleus (cf. the method of modified plane waves); however, the choice of the coefficient of this additional function can only be made by introducing additional hypotheses.

The orthogonalized plane wave method gives of course best results with metals. After the first application by Herring and Hill (1940) to Be, Callaway and co-workers

have made extensive applications to Fe (Callaway 1955 a, b), to K (Callaway 1956), to Na (Callaway 1958 b), to Cs (Callaway and Haase 1957), to Li (Glasser and Callaway 1958, Callaway and Glasser 1958). Heine (1957) has treated Al in a calculation in which a self-consistent field has most nearly been achieved. Fukuchi (1956) has treated Cu. Application to semiconductors and insulators has been fairly successful, but it implies a considerable effort. Whereas in a metal quite a small number of orthogonalized plane waves is sufficient, for a semiconductor one may require a large number of terms. The first application to non-metals was that of Herman (1952) to diamond. Then Ge was extensively studied by Herman and Callaway (1953) and Herman (1954). Other calculations are those of Woodruff (1955, 1956) and of Bassani (1957, 1959 b) on Si and that of Kobayasi (1958) on SiC. Bassani (1959 a) studied also a hypothetical Na with the Si structure, to test the importance of the potential in determining the band structure. In agreement with current ideas, it was found that sodium with the diamond structure would have a band structure very similar to that of Si, that is, the potential has not much effect. The same result was obtained with Li, instead of Na (Bassani and Celli 1959). In other cases, however, certain details of the band structure have been found to depend on the potential, for instance the relative position of the s-like and p-like levels at the centre of the dodecahedral faces of the Brillouin zone for body-centred alkali metals (cf. also Callaway 1959). Incidentally, an interesting discussion on the way in which many electronic properties can be correlated to the size of the band gap at these points has been given by Cohen and Heine (1958).

The inverse problem, that is the study of the importance of the crystal structure, was considered by Casella (1958), who carried out an orthogonalized plane wave calculation on a hypothetical face-centred cubic carbon. It was found that this substance would be an almost perfect metal, pointing out again the overwhelming importance of the crystal structure.

The possibility of using orthogonalized plane wave calculations to work out the matrix elements of momentum, and thus the optical absorption, has been examined by Suffczynski (1958). The use of the method as an interpolation procedure along the same lines as tight binding has been discussed by Allen (1955).

4.3.3. *Modified plane waves* (Brown and Krumhansl 1958).

The idea of this method is to introduce an auxiliary function ϕ, which is an approximate eigenfunction and may be found, for instance by a cellular calculation, to add to the plane waves. Considering the Fourier coefficients of this function ϕ, some of them (the lowest) will correspond to the plane waves used in the expansion of the wave function; the higher coefficients, however, will represent the high frequency oscillations of the wave function near the nucleus. If one subtracts the lowest plane waves from ϕ, the resulting function and the plane waves form a set of orthogonal functions which may be used as basis for a variational procedure.

The method can easily be extended to incorporate any number of auxiliary functions, and it includes the orthogonalized plane wave procedure as a particular case, when the auxiliary functions are the core atomic orbitals; however it avoids the difficulties mentioned in connection with orthogonalized plane waves. It has been applied so far only to Li, with satisfactory results.

4.4. *Methods using a Constant Potential outside Atomic Spheres* (Slater 1937)

In the cellular method one uses the expansion

$$\psi = \sum_{l,m} C_{l,m} P_l^m(\cos\theta) R_l(r) \qquad \ldots\ldots(21)$$

which is a good approximation near the nucleus, but requires many terms near the boundary of the Wigner–Seitz cell. In plane wave methods one uses the expansion

$$\psi = \sum_{\mathbf{K}_n} C_{\mathbf{K}_n} \exp[i(\mathbf{k}+\mathbf{K}_n).\mathbf{r}] \qquad \ldots\ldots(22)$$

which has properties opposite to those of (21). Slater (1937) suggested using (21) and (22) only in those regions in which they are expected to give a quickly convergent representation of the wave function. Take then non-overlapping spheres surrounding the atoms. Within each sphere the potential is assumed spherically symmetric and (21) holds. In the regions between the spheres the potential is assumed constant and (22) holds. For each plane wave one can choose the $C_{l,m}$ in (21) so that, carrying the summation to $l\to\infty$, the series (21) is continuous with the plane wave at all points on the surface of the sphere. There remains though a discontinuity in grad ψ at the surface, and this will give a contribution to the kinetic energy when the mean value of the Hamiltonian is calculated. The combination of the two solutions holding in the two regions is now called an augmented plane wave.

In the usual way one constructs trial functions as sums of these composite functions and applies the variational procedure. The method did not receive much application in its original form (Chodorow 1939) but was later modified, becoming the method of augmented plane waves (§ 4.4.1). There is also a modification (Leigh 1956, 1958) in which the fact is taken into account that in practice the summation over l in (21) cannot be taken to ∞, and thus ψ will not be continuous on the surface of the atomic sphere. Leigh then suggests taking explicit account of this fact, and deriving a secular equation from a variational procedure, for instance Kohn's procedure, which minimizes the mismatch on that surface effecting also the best possible compromise between a mismatch of ψ and a mismatch of grad ψ. The possible advantages of this method have not yet been fully explored.

Another modification is due to Matyáš (1952), who applied it to Be. The same procedure was applied by Trlifaj (1952) to Mg and by Antončík (1953) to Al.

4.4.1. *Augmented plane waves* (Slater 1953, Saffren and Slater 1953).

This is another way of matching solutions of the form (21) inside the atomic sphere to solutions of the form (22) outside it, and here one writes the plane wave as a superposition of a number of scattered spherical waves outside the atomic sphere (cf. § 4.4.2). As is well known, the plane wave $\psi_0 = \exp(i\mathbf{k}.\mathbf{r})$ can be expanded in a series of Legendre functions based on the atom under consideration

$$\psi_0 = \sum_l i^l(2l+1) P_l(\cos\theta) j_l(kr) \qquad \ldots\ldots(23)$$

the angle θ being measured relative to $\mathbf{k}$. In practice the summation extends over as many values of l as one can deal with.

Choosing now a trial value of the energy E_l, one finds the radial solutions of Schrödinger's equation inside the sphere, for each l, and constructs a linear combination of these solutions determining the coefficients so as to join continuously to (23) on the surface of the atomic sphere. Again there will be a discontinuity in grad ψ. The expectation value of the energy $\bar{E}$ is calculated for the augmented plane wave so constructed. There will be contributions to $\bar{E}$ from the region inside the sphere, from the region outside, and from the discontinuity in grad ψ at the surface. $\bar{E}$ will be a function of E_l and one varies E_l to make $\bar{E}$ stationary, assuming that such a stationary value is the best approximation to the correct energy. A pleasing feature of the method is that one finds that the value of E_l that makes $\bar{E}$ stationary is equal to the stationary value $\bar{E}$ itself. Thus an augmented plane wave yields an $\bar{E}$ such that the function is within each sphere an eigenfunction of Schrödinger's equation belonging to the eigenvalue $\bar{E}$. It is not an exact solution outside the sphere, because $\bar{E}$ is not generally equal to k^2, but

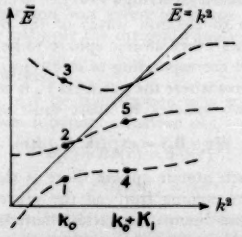


Figure 5. $(\bar{E}, k^2)$ curves in the method of augmented plane waves.

the error due to this fact is compensated by the contribution of the discontinuity in grad ψ. Another fortunate circumstance is that the relation between $\bar{E}$ and $\mathbf{k}$, the wave vector of the plane wave, can be expressed analytically : if Ω is the volume of the Wigner–Seitz cell minus the volume of the atomic sphere of radius r_0, then

$$\Omega(\bar{E}-k^2) = 4\pi \sum_l r_0^2(2l+1)\,j_l^2(kr_0)\left[\frac{d}{dr}\ln R_l(\bar{E},r)\right]_{r=r_0} . \qquad \ldots\ldots(24)$$

Note that it is required to know only the logarithmic derivative of the radial functions at $r = r_0$. The form of the curves one finds is illustrated schematically in figure 5. The various curves correspond to solutions within the sphere with different predominant values of l and different numbers of nodes. If the augmented plane waves were exact solutions, $\bar{E}$ would equal k^2, thus the solution to take as a first approximation is that which lies closer to the line $\bar{E} = k^2$. The various functions corresponding to the same $\mathbf{k}$ are orthogonal, so that here orthogonalization to the core functions is automatic. Plotting the same curves using the reduced zone scheme, one obtains an approximate band structure, spherically symmetric in $\mathbf{k}$-space, which, particularly if combined with the free electron picture and with results of simple tight binding computations, gives many hints on the true structure.

The next step is to take the augmented plane waves as basic functions for the variational procedure. Given $\mathbf{k}$, e.g. $\mathbf{k}_0$, one combines functions with different

E (points 1, 2, 3) and also functions whose plane wave part has vectors $\mathbf{k}_0 + \mathbf{K}_n$ (points 4, 5). One increases the number of augmented plane waves until the addition of further waves has no effect on the value of the energy to the required approximation. As expected since one employs in each region the most natural set of basic functions, the convergence is generally found to be good. The method thus appears very satisfactory. One starts of course with a very particular form of potential, but in the final variational calculation this restriction can be removed.

The method has been tested on several substances, however there is no published work which is completely satisfactory. Most extensive are Howarth's (1955) calculation on Cu. The method was extended by Saint-James (1958) to include spin–orbit interaction. Note, that since the periodicity of the lattice is not used in the first part of the augmented plane wave procedure, this can be applied to disordered substances, e.g. to liquids.

4.4.2. *Scattering matrix method* (Korringa 1947).

This is an application to electron waves of the dynamical theory of lattice interferences. One assumes every atomic sphere to be the source of an outgoing wave of wave vector $\mathbf{k}$ and corresponding to electrons of energy E propagating in the region outside the spheres where the potential V_0 is constant. Let $k_0^2 = E - V_0$. The waves emitted by two spheres $\mathbf{R}_i$ apart must satisfy the phase condition

$$\phi(\mathbf{r} + \mathbf{R}_i) = \exp(i\mathbf{k} \cdot \mathbf{R}_i)\,\phi(\mathbf{r}).$$

Now, at the surface of each atomic sphere there is the wave outgoing from the sphere, ϕ_{out}, and the waves coming from all the other spheres ϕ_{in}, and the two must be the same. Expand them in spherical harmonics, the coefficients being respectively a_{lm} and b_{lm}. Now ϕ_{in} can be expressed in terms of ϕ_{out} through the phase condition ; and k_0, that is the energy, enters into the resulting equations for a_{lm}, b_{lm} as an unknown parameter. On the other hand ϕ_{out} must be related to ϕ_{in} as scattered and incident waves, that is through a scattering matrix depending on the potential $V(r)$ inside the atomic sphere. If $V(r)$ has spherical symmetry, the usual scattering theory gives

$$a_{lm} = \tfrac{1}{2} b_{lm}[1 - \exp(-2i\eta_l)],$$

η_l being a phase shift which is a function of the energy and is determined by the condition that the solution $K(r)$ of

$$\frac{d^2K}{dr^2} + \left[k_0^2 - V(r) + V_0 - \frac{l(l+1)}{r^2}\right] K = 0$$

outside the sphere must have the form

$$K(r) = C \sin(k_0 r - \tfrac{1}{2}\pi l + \eta_l).$$

If the potential is not spherically symmetric, each a_{lm} is no longer determined only by the corresponding b_{lm}, but is a linear function of all the b_{lm}, the coefficients being the elements of the scattering matrix. In any case one has two sets of equations for the coefficients a_{lm}, b_{lm}, which can be satisfied simultaneously only for certain values of k_0, that is of E. $\mathbf{k}$ enters into the

first set of equations, and thus the relation $E(\mathbf{k})$ is determined. The first part of the calculation, that is the determination of ϕ_{in} as the superposition of the waves coming from all the other spheres, requires the calculation of lattice sums for many values of $\mathbf{k}$, but does not depend on the potential and thus, apart from scaling factors, is the same for all crystals with the same structure and can be carried out once and for all for any given structure, an advantage shared by many methods based on a constant potential outside atomic spheres. The second part, the scattering problem, is independent of $\mathbf{k}$. In practice it can be carried out only for a spherically symmetric potential.

This method is physically very interesting. In its original form it was applied only to Nb (Trefftz 1954). However, the variation–iteration procedure (§ 4.4.3) leads, when one assumes a constant potential outside atomic spheres, as one must do in practice, to the same formalism, and many applications can be anticipated. There are the same advantages as with augmented plane waves, and in addition the difficulty of a discontinuity in the slope is avoided. An instructive discussion of the method in one dimension was given by Saxon and Hunter (1949), who also discussed on the same example the integral equation approach.

4.4.3. *Integral equation approach.*

Schrödinger's equation may be transformed into an integral equation by the general method of the Green function. Instead of

$$\nabla^2\psi(\mathbf{r}) + E\psi(\mathbf{r}) = V\psi(\mathbf{r}) \qquad \ldots\ldots(25)$$

consider

$$\nabla^2 G(\mathbf{r},\mathbf{r}') + EG(\mathbf{r},\mathbf{r}') = \delta(\mathbf{r}-\mathbf{r}') \qquad \ldots\ldots(26)$$

where G is the Green function satisfying the same boundary conditions as ψ. It may be verified that G may be written, for instance, in the form

$$G(\mathbf{r},\mathbf{r}') = -\frac{1}{\Omega}\sum_{\mathbf{K}_n}\frac{\exp\left[i(\mathbf{k}+\mathbf{K}_n).(\mathbf{r}-\mathbf{r}')\right]}{|\mathbf{k}+\mathbf{K}_n|^2 - E}$$

and has the property $G^*(\mathbf{r},\mathbf{r}') = G(\mathbf{r}',\mathbf{r})$. Multiplying (25) by G^* and the complex conjugate of (26) by ψ, subtracting and integrating with respect to $\mathbf{r}$ over unit cell, the

$$\int (G^*\nabla^2\psi - \psi\nabla^2 G^*)\, d\tau$$

may be transformed into a surface integral which vanishes because of the boundary conditions, and one is left with

$$\psi(\mathbf{r}') = \int_{\text{unit cell}} G^*(\mathbf{r},\mathbf{r}')\, V(\mathbf{r})\,\psi(\mathbf{r})\, d\mathbf{r}$$

equivalent to

$$\psi(\mathbf{r}) = \int_{\text{unit cell}} G(\mathbf{r},\mathbf{r}')\, V(\mathbf{r}')\,\psi(\mathbf{r}')\, d\mathbf{r}' \qquad \ldots\ldots(27)$$

which is the integral equation satisfied by ψ. If instead of $\psi(\mathbf{r})$ one uses its Fourier transform, that is the probability amplitude of momentum $\psi(\mathbf{p})$, (27) may be written in the equivalent form

$$(p^2 - E)\,\psi(\mathbf{p}) = \int V(\mathbf{p}-\mathbf{p}')\,\psi(\mathbf{p}')\, d\mathbf{p}' \qquad \ldots\ldots(28)$$

where $V(\mathbf{p}-\mathbf{p}')$ is the Fourier transform of the potential, reducing in the periodic problem to a sum of δ-functions with coefficients $\phi(\mathbf{K}_n)$. The form (28) is probably more convenient in the general case for a solution by successive approximations (iterations). A succession of iterations and variations (Dank and Callen 1952) would lead most rapidly to a sufficient approximation to the eigenvalue. This method has not however been adequately tested. In any case the integral equation (27) or (28) is difficult of solution without restrictions on the potential, but Kohn and Rostoker (1954) have shown that one achieves a great simplification with the approximation of a constant potential outside atomic spheres, because then the integration in (27) is to be carried out only within the sphere (outside one can take $V = 0$ without loss of generality), where moreover V is spherically symmetric. The procedure then is to substitute to the integral equation a variational principle, using as trial function the usual expansion (21) in spherical harmonics. One has to expand also the Green function in spherical harmonics and, as already mentioned, the formalism becomes identical to that of the S-matrix approach, the main difficulty being the evaluation of complicated lattice sums. The method has to be tested only on Li (Kohn and Rostoker 1954), the convergence being found to be good, and preliminary calculations for crystals with the diamond structure, mainly in view of application to Ge, have been reported by Segall (1959), who had previously (1957) extended the method to polyatomic crystals. The Green function formalism has been applied to a simple model by Engelmann (1956) and by Raychaudhuri (1957).

4.4.4. *Use of a constant potential outside atomic spheres in conjunction with the cellular procedure.*

The main advantage of a constant potential outside atomic spheres, that is the fact that a large part of the calculation is the same for all crystals with the same structure, can be exploited also within the framework of the cellular method (Pincherle, unpublished). Again, one uses the expansion (21) inside the atomic sphere, while in the region between the sphere and the boundary of the Wigner–Seitz polyhedron one uses the expansion

$$\psi = \sum_{l,m} C_{lm} P_l^m(\cos\theta)\,[j_l(E^{1/2}r) + b_l n_l(E^{1/2}r)] \qquad \ldots\ldots(29)$$

where the C_{lm} are determined as usual from symmetry (apart from a normalization factor), j_l and n_l are the two linearly independent spherical Bessel functions of order l and b_l is determined by the condition that the logarithmic derivative of any radial function in (29) should be equal to that of the corresponding function $R_l(r)$ in (21) at $r = r_0$. Any particular crystal of the given structure is then characterized by the logarithmic derivatives of the R_l functions at $r = r_0$. These must be known for a range of values of the trial energy E_t and often present the advantage of being smooth functions of E_t so that they can be interpolated from a knowledge of a few values (cf. the quantum defect method). The part of the calculation common to all crystals with the same structure is the matching of the solution in one cell to those in adjacent cells, and for this purpose one may employ any of the methods described in § 4.1. One obtains the elements of a determinantal equation expressed as functions of $E^{1/2}r_0$ and of the logarithmic derivatives of the

R_l's at $r = r_0$. Suitable tables may be constructed; then any special case will entail only the solution of secular equations.

The method has been tested only on simple examples. It is probably easier to apply than other methods making use of a constant potential outside atomic spheres. On the other hand, there is still the not completely resolved question of the boundary conditions in the cellular procedure, and deviations of the potential from the simple form assumed are more difficult to take into consideration.

4.5. *Methods using Empirical Parameters*

4.5.1. *The quantum defect method* (Kuhn and Van Vleck 1950, Brooks 1953).

Mainly devised for calculation of cohesive energies, it can be useful also in band structure calculations as a short cut. An extensive review has been given by Ham (1955). It starts from the assumption that, for valence and conduction bands, the one electron Hamiltonian in a crystal is, near each atom, the same as for a valence electron in a free atom, and uses the fact that it is possible to construct the eigenfunctions of this Hamiltonian, for distances larger that an effective radius r_0, from the observed spectroscopic term values of the free atom. Since in many methods, particularly in the cellular method, the $(E, \mathbf{k})$ relation is determined only from these solutions outside the core, it follows that it is not necessary to know the potential within it, and many numerical calculations are eliminated.

For the wave functions outside the core one could take Bessel functions, as in §4.4.4, but most calculations with the quantum defect method have been made assuming that they may be taken as Coulomb eigenfunctions, that is solutions of the radial Schrödinger equation with a Coulomb potential. There are two independent functions for each value of l and the appropriate linear combination is again determined by the logarithmic derivative D_l of the radial function at $r = r_0$. If one knows the D_l's for all energies, the required eigenfunctions are determined. Now it has already been remarked that the D_l's are reasonably smooth functions of the energy at not too large r_0. When the energy is an atomic eigenvalue, the radial function vanishes at ∞, and from this one calculates the corresponding D_l (r_0), provided r_0 is outside the region where the potential deviates from the Coulomb form. The D_l's for any energy are obtained by interpolating between the eigenvalues.

This was the original idea, but it was found that the interpolation is not always reliable so that one should look for another quantity which is a smoother function of the energy than D_l. Such a quantity is the quantum defect δ_m which appears in Rayleigh's expression for the energy ϵ_m of an atomic level $\epsilon_m = -(m-\delta_m)^{-2}$ where m is a whole number which increases by one between adjacent terms of the same series. The δ's are different for different series, but within one series they vary little and very smoothly. Now it is possible to show that, with the usual choice of the Coulomb eigenfunctions, the ratio β_l of their coefficients at the eigenvalue is given with good approximation by

$$\beta_l = -\frac{\Gamma(n+l+1)}{n^{2l+1}\,\Gamma(n-l)\tan\pi\delta_m}$$

where $n^{-2} = E$ and δ_m is the experimental quantum defect. It is then assumed that the same expression holds at intermediate energies, with the interpolated

value of δ. Of course, knowing the wave functions for $r \geqslant r_0$, the logarithmic derivatives at r_0 can also be calculated.

The method has been further elaborated and refined by Brooks and Ham (1958), who showed that the amplitude of the wave function near the nucleus also can be derived from the quantum defects, and argued that the quantum defect method takes very general account of exchange and correlation interactions between the valence and the core electrons. Relativistic effects, including spin–orbit coupling, are also included naturally. Corrections were also worked out to take into account the polarization of the core and deviations of the potential from the Coulomb form. Also Callaway, Woods and Sirounian (1957) used the quantum defect method to discuss relativistic effects.

4.5.2. *The* $\mathbf{k}.\mathbf{p}$ *approximation.*

This is a powerful tool in the problem of deducing information on band structure directly from experiment. It was first used (Seitz 1935, Bardeen 1938) as a simple perturbation calculation. In Schrödinger's equation

$$\mathscr{H}\psi_{n\mathbf{k}} = E_{n\mathbf{k}}\psi_{n\mathbf{k}}$$

write $\psi_{n\mathbf{k}} = \exp(i\mathbf{k}.\mathbf{r})\,u_{n\mathbf{k}}(\mathbf{r})$, where $u_{n\mathbf{k}}$ has the periodicity of the lattice; substituting one finds the equation for u

$$[\mathscr{H} + \mathbf{k}.\mathbf{p} + k^2]\,u_{n\mathbf{k}} = E_{n\mathbf{k}}u_{n\mathbf{k}}$$

and, if u_{n0} is known, one can treat the term $\mathbf{k}.\mathbf{p}$ as a perturbation, obtaining

$$E_n(\mathbf{k}) = E_n(0) + k^2 + \sum_{n \neq n'} \frac{(\mathbf{k}.\mathbf{p})_{nn'}(\mathbf{k}.\mathbf{p})_{n'n}}{E_n - E_{n'}} + \ldots \qquad \ldots\ldots(30)$$

A similar expression may be obtained also when the edge is not at $k = 0$. When the bands are degenerate the expression is more complicated (Shockley 1950). In any case, the important point is that $E_n(\mathbf{k})$ may be expressed by means of parameters, chiefly elements of the effective mass tensor, which it is not necessary to calculate because they can be derived from experiment, e.g. from cyclotron resonance observations.

The perturbation treatment is limited to a small region of $\mathbf{k}$-space; when the perturbing term is of the order of the separation of the bands at $k = 0$, one must treat the interaction between adjacent bands exactly. This can be done in principle, and formulae with a number of parameters obtained, but the number of these unknown parameters increases rapidly with the number of bands, so that many accurate experimental results are necessary before one can draw reliable conclusions. Further complications arise when one wishes to take into account also spin–orbit coupling.

Kane (1956, 1957, 1959) has applied this approach to the determination of the structure near the edge of the valence band of Si and Ge, and of the valence and conduction band of InSb, with an attempt to cover all the bands for Ge. Much work still remains to be done to refine the method; for instance, to take exchange into account, the matrix elements of $\mathbf{p}$ must be multiplied by $1 + \epsilon$, but ϵ cannot be determined with any accuracy. Nevertheless this appears one of the most reliable ways of deriving band structures from experimental data.

4.5.3. *Pseudo-potentials* (Phillips 1958).

This is an interpolation scheme, of the type of those based on tight binding or on orthogonalized plane waves, but probably simpler to use and more effective. It exploits experimental data, such as direct and indirect energy gaps, and locations of edges, besides data on the effective mass tensor, and is based on the nearly free electron formalism so that it may be considered as a semi-empirical alternative to orthogonalized plane waves. It splits up the potential for electrons near the Fermi surface into two parts, one due to the core, the other to the valence electrons, and assumes that for semiconductors the relative effect of the core is small, at least on the variation of the energy with $\mathbf{k}$ near the band edges. It has in fact been found in many calculations that the high Fourier coefficients of the potential are not important in determining effective masses, even if they affect the relative position of the levels. One uses then only the lowest Fourier coefficients of the potential in the nearly free electrons formalism (§ 4.3.1), but determines them empirically, in such a way as to fit the above mentioned experimental data. This is equivalent to using an effective repulsive potential due to the core electrons, which replaces the orthogonalization terms of orthogonalized plane waves. Phillips constructed in this way three parameters 'pseudo-potentials' for Si and Ge which are claimed to yield fairly accurate ($\delta E < c.\,0{\cdot}7$ ev) sketches of the energy bands of these crystals in the main directions of k-space near the band edge. Also this approach will have to be further refined. The use of a repulsive potential in substitution for the orthogonality condition in the orthogonalized plane waves method has been discussed by Antončík (1959): it seems to simplify calculations without too great a loss of accuracy. A rigorous formulation of the empirical potential approach has been given recently by Phillips and Kleinman (1959): this discussion clarifies also many aspects of the orthogonalized plane waves procedure.

§ 5. The Construction of the Crystal Potential

This part of the problem has attracted less attention, even though it was thought for a time that the characteristics of the potential were all important in determining a band structure. To-day it begins to appear that they are not in fact so important (see § 4.3.2, and also Fues and Statz (1952)); moreover, the tendency is to determine band structures without having to introduce a potential explicitly. Nevertheless, for the purpose of deriving results exclusively from first principles, it is very desirable to put the choice of the potential on a firm foundation.

In dealing with this problem the point of view has always been so far that of the self-consistent field, but many different simplifying assumptions have been used. It is clear also that the approximations must be different in an ionic crystal, in a valence crystal or in a metal; but in no case has a generally accepted approach developed yet.

As is well known, adopting the point of view of Hartree's self-consistent field, each electron is assumed to move in the field $\mathbf{F}$ produced by the nuclei and by the charge density of all the other electrons, the latter being obtained from one-electron wave functions $u_i(\mathbf{x}, \sigma)$ ($\mathbf{x}$ denotes the space coordinates and σ the spin coordinate), which are solutions of Schrödinger's equation for the field $\mathbf{F}$. In the Hartree–Fock approach one takes into account the Pauli principle and writes ψ as a completely antisymmetrical combination of the $u_i(\mathbf{x}_j, \sigma_j)$. In this way, exchange

is taken into account, but one still neglects configuration interaction, whose consideration requires the use of more than one of such determinantal combination. For the Hamiltonian one writes

$$\mathscr{H} = \mathscr{H}_0 + \sum_i f_i + \sum_{i,j} g_{ij}. \qquad \ldots\ldots(31)$$

Here $\mathscr{H}_0$ is independent of the coordinates of the electrons; f_i depends on the coordinates of only one electron and includes its kinetic energy, its potential energy in the field of the nuclei and external fields, and also spin–orbit interaction and other minor terms—generally neglected; g_{ij} depends on the coordinates of two electrons and is in practice their Coulomb repulsion, magnetic interactions being usually neglected. With these assumptions one obtains the Hartree–Fock equations, e.g. for electron $\mathbf{x}_1$ in state u_i (the u_i indicate now only the space part of the one-electron functions)

$$f_1 u_i(\mathbf{x}_1) + \left[\int \sum_j u_j^*(\mathbf{x}_2) u_j(\mathbf{x}_2) g_{12}\, d\mathbf{x}_2\right] u_i(\mathbf{x}_1) - \sum_{\substack{j \\ \text{spin } j = \\ \text{spin } i}} \left[\int u_j^*(\mathbf{x}_2) u_i(\mathbf{x}_2) g_{12}\, d\mathbf{x}_2\right] u_j(\mathbf{x}_1) = \epsilon_i u_i(\mathbf{x}_1). \qquad \ldots\ldots(32)$$

All the summations are taken over the N occupied orbitals, but there are solutions corresponding also to excited unoccupied states. The eigenvalue ϵ_i is to a good approximation the negative of the energy required to remove an electron from state i. The second term in (32) is the potential energy of the electron in state u_i at $\mathbf{x}_1$ due to the charge distribution of all the N electrons, including electron 1 itself. The first two terms give thus the classical picture of the self-consistent field, but there is included also the interaction of the electron with itself. If the third term is not considered, as in the original Hartree treatment, one must then subtract the term

$$\left[\int u_i^*(\mathbf{x}_2) u_i(\mathbf{x}_2) g_{12}\, d\mathbf{x}_2\right] u_i(\mathbf{x}_1); \qquad \ldots\ldots(33)$$

in other words, one must remove from the potential the contribution of the charge distribution of the electron under consideration; this is what is done in calculations on atoms. A similar procedure was adopted also in the early calculations on solids by the cellular method. The subtraction however does not appear necessary when calculating the potential acting on an electron in a normally unoccupied orbital, because then the self-energy is not included, the summation in (32) being only over the occupied states. This point, to be further discussed below, is one of the difficulties in setting up a potential for a solid. Another difficulty is the question whether the potential can be taken as the sum of potentials of free atoms or ions: of course in a crystal the Hartree procedure taken literally is not satisfactory because the u_i extend through the whole crystal, and thus near each atom the correction (33) is negligible. This means that near each atom the field approaches that of a neutral atom. Now physical intuition tells one that this may be right for additional electrons not belonging to the atom itself (conduction band of a semiconductor), but cannot be correct for an atomic electron (valence band), as may be seen by considering the case of widely separated atoms. Further complications arise in

metals and ionic crystals. A discussion of these questions, from the point of view of the equivalent orbitals picture, has been given by Hall (1956), his main result being that the crystal potential in most crystals should be quite close to the superposition of atomic or ionic fields; the potential must be different, in the sense discussed above, for occupied and unoccupied states, and moreover is not constant even over the states of an occupied or an unoccupied band.

To gain a physical picture, consideration of the third term, the exchange term, in (32) is helpful. It is more complicated than the Coulomb term since the function u_i to be determined appears under the integral sign. Slater (1951 b) has shown that it can however be written formally as the second term, namely as

$$\left[\int \rho(\mathbf{x}_2, \mathbf{x}_1) g_{12}\, d\mathbf{x}_2\right] u_i(\mathbf{x}_1) \qquad \ldots\ldots(34)$$

by introducing a suitable 'exchange charge density' $\rho(\mathbf{x}_2, \mathbf{x}_1)$. This exchange charge density differs however for each solution of the Hartree–Fock equations.

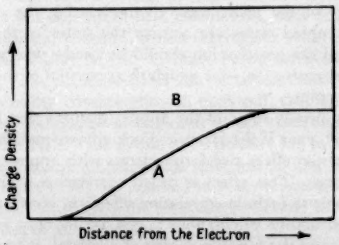


Figure 6. Variation with interelectronic distance of the exchange charge density. A: electrons with same spin; B: electrons with opposite spin.

To make the concept useful it is then necessary to form a weighted average over all occupied orbitals of the same spin. The use of density matrices (Maslen 1956) yields a definition of the exchange hole which is identified to Slater's averaged quantity. The integral (over, say, $\mathbf{x}_2$) of this exchange charge density is equal to one; since the exchange term appears with a minus sign in (32), it provides the required subtraction of one negative charge. This however holds only if u_i is an occupied orbital. If it is an excited state, the total exchange charge vanishes. This agrees with the conclusion reached above from an intuitive point of view on when one should remove one negative charge. There remains the problem of where this charge should be removed from: not, as subtraction of (33) would imply, almost uniformly from the whole of the crystal; rather, a common approximation was to remove it from an appropriate atomic state, for instance, when dealing with the valence band of silicon, one would remove it from the 3p atomic level in each cell, or perhaps partly from the 3p and partly from the 3s level. However, when one examines how the exchange charge density varies with the distance from a given electron, one finds (see, for instance Slater 1956, page 52) that it varies as indicated in figure 6. Curve A refers to electrons with the same spin as

the electron in question (as is well known, electrons of the same spin keep away one from the other as an effect of the Pauli principle), and curve B to electrons with opposite spins. The picture that emerges is of a negative charge removed not from an atom or from an electronic state, but from around the electron under examination. The electron is thus surrounded by a positive hole (exchange hole), whose shape depends both on the state of the electron in question, and on its position. Again this indicates that one should use different potentials for the different electrons.

A distinction should here be drawn between methods using a potential which is made up of atomic-like potentials centred on the various atoms, and the plane wave methods using a crystal potential expressed, say, through its Fourier coefficients (observe that such a potential can be constructed out of atomic potentials in an infinite number of ways). In this last case it is much more plausible to argue that the same potential can be used for both occupied and excited states, since each wave function extends over the whole crystal and hence its effect on the potential at any point is negligible. Indeed, this may be considered a practical advantage of the plane wave methods. On the other hand, with an atomic approach, occupied states are akin to the states of the neutral atom—for which the potential of the positive ion should be used—whereas excited states are akin to states of the negative ion—for which the potential of the neutral atom is the appropriate starting point.

Before examining briefly some of the approximations that have been used, one should remember that, even if the Hartree–Fock approximation does not include it, there must be a repulsion effect also for electrons with opposite spins, just because of their electric charge. This effect is called correlation. Until a few years ago it seemed impracticable to include correlation effects in constructing a one-electron potential for band structure calculations: one could only include a correlation energy when working out the binding energy of a crystal, following an approximate treatment of Wigner and Seitz (1934). Physically, one has for the electrons of opposite spin a correlation hole, qualitatively similar to the exchange hole holding for the electrons with the same spin. Only the consideration of configuration interaction (the use of several determinantal wave functions corresponding to different electronic configurations) would include this correction in the Hartree–Fock method.

To-day, however, the two effects can be included together if one uses the treatment of electron interactions introduced by Bohm and Pines (see for instance Pines 1955). As developed so far, the theory applies to conduction electrons in metals, which are almost free. The effect of the crystal field has been investigated (Hubbard 1954), but no complete treatment is available yet. For our purpose the main result is that the field of each electron is screened by the presence of the other electrons so that it has a potential of the form

$$e/r \exp(-ar). \qquad \ldots\ldots(35)$$

In metals a^{-1} is of the order of the lattice distance.† Therefore the field of any electron is felt only up to distances of this order of magnitude: for larger distances

† As is well known, the treatment of interactions by means of a screened Coulomb potential is valid only at small distances. The motion due to long range interactions is better analysed in terms of a set of plane waves—the plasma oscillations.

the charge has been completely compensated and again one may express this by saying that the electron is surrounded by a hole, whose total charge is $+1$, of radius approximately a^{-1}. There is no distinction here between electrons of opposite spins, therefore this hole is a combination of the exchange and correlation holes. It is a simplification to be able to treat the two effects at the same time. The other great advantage of this approach is that in first approximation the hole manifests itself in a contribution to the potential which has a very simple form, being dependent on $\mathbf{k}$, but constant in space. It thus only adds to the energy a quantity $V_{ec}(\mathbf{k})$ that can be written down explicitly as a function of $\mathbf{k}$ (see Pines 1955, page 407). In this way it has become practicable, at least for metals, to carry out self-consistent calculations (Heine 1957, on Al).

5.1. *The Approximations Most Widely Used*

For monatomic substances, in most calculations based on atomic orbitals or cellular methods, the approach used originally by Wigner and Seitz (1933) on Na has been followed. The idea is to use the field for the isolated atom. This may be known through a self-consistent field calculation, with or without exchange, or it may be a semi-empirical potential which reproduces some of the energy levels of the isolated atom, or it may be the Thomas–Fermi statistical potential (Gombás 1949, March 1957). For occupied states one substracts from this field the contribution of the electron under consideration in each cell, that is one assumes that the exchange hole is just the charge distribution of the electron itself within one cell, not extending to other cells. Thus one has a positive ion in each cell, with neutral atoms outside : these atoms are supposed spherical so that they produce no field. The discussion by Hall (1956) has justified in some respects this approach.

Other authors have considered the total charge density in the crystal as the sum of the charge densities of all the atoms

$$\rho_{\text{crystal}}(\mathbf{r}) = \sum_{\mathbf{R}_i} \rho_{\text{atom}}(\mathbf{r} - \mathbf{R}_i) \qquad \ldots\ldots(36)$$

each charge density extending generally beyond the boundary of the corresponding cell. With respect to the previous assumption, this has the effect of lowering the interatomic potential barriers. Expanding the periodic function $\rho(\mathbf{r})$,

$$\rho(\mathbf{r}) = \sum_{n} \rho(\mathbf{K}_n) \exp(i\mathbf{K}_n \cdot \mathbf{r}),$$

the corresponding potential has the form

$$U(\mathbf{r}) = \text{constant} - 8\pi \sum_{n}{}' \frac{\rho(\mathbf{K}_n)}{K_n^2} \exp(i\mathbf{K}_n \cdot \mathbf{r})$$

(the prime indicating that $K_n = 0$ is excluded from the summation), particularly suitable for calculations with plane wave methods, since it is expressed through its Fourier coefficients. If one wants explicitly the potential in each cell, one averages spherically in the cell the charge densities arranged on the various atomic cells and normalizes to the correct number of electrons, and then one obtains the potential of the spherical distribution as for an atom.

For taking into account the exchange hole, a simplified method of Slater (1951 b) has been widely used : the exchange charge, already averaged over all occupied

orbitals, is taken to be spherically symmetric around each electron and determined by the local charge density of electrons with the same spin, which is assumed to be the same as for a free electron gas of the same overall density. The adequacy of this approximation, and an alternative way of dealing with the exchange correction, have been discussed by Herman, Callaway and Acton (1954). For the case of iron, where a fairly accurate calculation was made by Callaway (1955 a), it was found that the Slater free electron simplification overestimates the exchange energy by about 15–20%.

Further complications arise in the case of polyatomic crystals. Consider for instance the case of lead sulphide. The valence band contains six electrons per pair of atoms, and these electrons are mainly concentrated in the sulphur cell. When calculating the potential acting on these electrons, one will have to remove one negative charge in all from the two cells, and a preliminary calculation will be necessary to estimate the total charge of the cells and how much should be removed from each. In such an ionic crystal one must also take into consideration the effect of all the ions in the lattice, which produce a ' Madelung potential ' different in the positively charged lead cell and in the negatively charged sulphur cell, and which depends on the total charge of the cells. Also, when one has removed some charge from, say the S cell, one has to consider the effect of this positive hole on the potential in the Pb cell. It is clear that all this involves a lot of more or less plausible guessing and that the situation is unsatisfactory (Bell *et al.* 1953, Birman 1958).

As already remarked, the Bohm–Pines theory makes possible a more firmly grounded treatment in the case of metals. As an example, we may discuss the potential used by Heine (1957) for his orthogonalized plane wave calculation on Al. The total effective potential acting on a conduction band electron was taken to be

$$U(\mathbf{r}, \mathbf{k}) = \text{constant} + V_{ec}(\mathbf{k}) + \sum_j V_1(|\mathbf{r} - \mathbf{R}_j|),$$

where $V_{ec}(\mathbf{k})$ is the exchange and correlation term given by the Bohm–Pines approach, and the spherically symmetrical potentials V_1, centred on the various atoms and considerably overlapping, are taken not to depend on $\mathbf{k}$. The contributions to V_1 taken into consideration were:

(1) The self-consistent potential with exchange calculated for the Al^{3+} core, slightly modified to include also configuration interaction.

(2) Exchange between inner core and conduction electrons, expressed through Slater's exchange charge density. Slater's free electron approximation, however, was not introduced, but it was assumed instead that the conduction electrons are either in a 3s or in a 3p state. One obtains thus one exchange potential for s-like states and one, slightly different, for p-like states. This is better than may appear at first sight, since the admixture of states with other values of l is not generally large. Perhaps more serious was the smoothing out necessary near the zeros of the 3s and 3p functions, where the exchange charge density fluctuates widely.

(3) The potential due to the conduction electrons inside an equivolume sphere. Here, in order to average in a simple way the charge distributions of the various states, ψ for these electrons was assumed to be one single orthogonalized plane wave, and the Fermi surface was taken to be just the sphere containing three electrons per

atom. Since the simple potential due to the spheres cannot be self-consistent, and also to account for the gaps between the overlapping spheres, a correction had to be applied, estimated from a simple model (a face-centred structure of protons in a uniform distribution of electrons, the total electric charge vanishing).

(4) Another small term taking into account the fact the V_{ec} does in fact depend slightly on position.

The potential so constructed could be made self-consistent without much trouble because the wave functions of the conduction electrons were found to differ little from just one orthogonalized plane wave, and thus the potential calculated from the true ψ's of these electrons differs little from contributions (3), thereby allowing self-consistency to be reached through a perturbation calculation.

The energy values of the unoccupied states were calculated with the same potential, which would seem to require some justification even if, as already remarked, this procedure is more plausible when dealing with a metal by plane wave methods.

This example shows that for metals the construction of the crystal potential can now be achieved in a fairly satisfactory manner. The situation is worse in the case of insulators and semiconductors, for which no satisfactory way of dealing with correlations is yet available, and for which there remains the difficulty of the different potentials that one should use for the valence and conduction bands.

§6. Concluding Remarks

Modern computational facilities have made practicable the majority of the many methods available for band structure calculation. In general, pushed to a sufficiently high approximation, all methods lead to the same results if they start from perfectly equivalent assumptions, and one will choose the method for which the initial assumptions seem most reasonable and which will lead to a convergent result with not too prohibitive labour. That the results of the various methods are close together has been shown by the extensive calculations by different methods carried out on alkali metals, particularly on lithium, whose band structure has been the subject of eleven investigations. Also in the more complex case of silicon, orthogonalized plane wave methods (Woodruff 1955, 1956, Bassani 1957) and cellular methods (Jenkins 1956) have led substantially to the same qualitative results, in broad agreement with deductions from experimental data.

One can thus say that generally the main features, down to considerable detail, of the band structure are yielded fairly satisfactorily by first principle calculations, even though full quantitative agreement with experiment is lacking. The main outstanding problems are the construction of satisfactory crystal potentials and achieving reasonable self-consistency.

There are also other checks on the results, because the wave functions obtained may be used to calculate theoretical values for other measurable quantities. Among these are the cohesion energy and the elastic constants, though this problem involves other factors and is outside the scope of the present review. For a qualitative analysis of cohesion in metals see Wigner and Seitz (1955). Another measurable quantity is the Knight shift (difference between the nuclear resonance frequency for an isolated atom and that for the same atom forming part of a crystal, which depends on the square of the wave function at the nucleus. For this

quantity, fairly good agreement has been obtained in some cases, e.g. for Rb (Callaway and Morgan 1958) and for Li and Na (Jones and Schiff 1954). But in general little has been done to obtain an evaluation of how good are the wave functions one has employed. Of course, particularly with variational procedures, fair agreement in the energy values does not guarantee that the wave functions will give correctly other physical quantities.

The calculations from first principles have their main usefulness in providing a complete, even if not yet very accurate picture of the energy bands. On the other hand the semi-empirical computations give an accurate picture of the structure, but only within limited regions. The two approaches are thus complementary and both necessary to gain a complete understanding of the electronic properties of crystals. The calculations of the first type have also great intrinsic interest as a challenge to devising new methods, to introducing the most reasonable simplifications, and to modern computational facilities.

References

Allen, L. C., 1955, *Phys. Rev.*, **98**, 993.

Altmann, S. L., 1958, *Proc. Roy. Soc.* A, **244**, 141, 153.

Altmann, S. L., and Cohan, N. V., 1958, *Proc. Phys. Soc.*, **71**, 383.

Antončík, E., 1953, *Czech. J. Phys.*, **2**, 18; 1959, *J. Phys. Chem. Solids*, **10**, 314.

Ariyama, K., and Mase, S., 1954, *Progr. Theor. Phys.*, **12**, 244.

Aubrey, J. E., and Chambers, R. G., 1957, *J. Phys. Chem. Solids*, **3**, 128.

Azbel', M. Ia., 1958, *Zh. Eksper. Teor. Fiz.*, **34**, 754.

Azbel', M. Ia., and Kaner, E. A., 1958, *J. Phys. Chem. Solids*, **6**, 113.

Bardeen, J., 1938, *J. Chem. Phys.*, **6**, 366.

Bassani, F., 1957, *Phys. Rev.*, **108**, 263; 1959 a, *J. Phys. Chem. Solids*, **8**, 375; 1959 b, *Nuovo Cim.*, **13**, 244.

Bassani, F., and Celli, V., 1959, *Nuovo Cim., Suppl.*, **11**, 438.

Behringer, R. C., 1959, *Phys. Rev.*, **113**, 787.

Bell, D. G., 1954, *Rev. Mod. Phys.*, **26**, 311.

Bell, D. G., Hum, D. M., Pincherle, L., Sciama, D. W., and Woodward, P. M., 1953, *Proc. Roy. Soc.* A, **217**, 71.

Bethe, H., 1929, *Ann. Phys., Lpz.*, **3**, 133.

Betts, D. D., 1959, *Canad. J. Phys.*, **37**, 350.

Bilz, H., 1958, *Z. Phys.*, **153**, 338.

Birman, J. L., 1958, *Phys. Rev.*, **109**, 810; 1959, *J. Phys. Chem. Solids*, **8**, 35.

Bloch, F., 1928, *Z. Phys.*, **52**, 555.

Bouckaert, L. P., Smoluchowski, R., and Wigner, E., 1936, *Phys. Rev.*, **50**, 58.

Brooks, H., 1953, *Phys. Rev.*, **91**, 1027.

Brooks, H., and Ham, F. S., 1958, *Phys. Rev.*, **112**, 344.

Brown, E., and Krumhansl, J. A., 1958, *Phys. Rev.*, **109**, 30.

Callaway, J., 1955 a, *Phys. Rev.*, **97**, 933; 1955 b, *Ibid.*, **99**, 500; 1956, *Ibid.*, **103**, 1219; 1957, *J. Electronics*, **2**, 330; 1958 a, *Solid State Phys.*, **7**, 99; 1958 b, *Phys. Rev.*, **112**, 322; 1958 c, *Ibid.*, **112**, 1061; 1959, *Ibid.*, **115**, 346.

Callaway, J., and Glasser, M. L., 1958, *Phys. Rev.*, **112**, 73.

Callaway, J., and Haase, E. L., 1957, *Phys. Rev.*, **108**, 217.

Callaway, J., and Morgan, D. F., 1958, *Phys. Rev.*, **112**, 334.

Callaway, J., Woods, R. D., and Sirounian, V., 1957, *Phys. Rev.*, **107**, 934.

Carlson, B. C., and Keller, J. M., 1957, *Phys. Rev.*, **105**, 102.

Casella, R. C., 1956, *Phys. Rev.*, **104**, 1260; 1958, *Ibid.*, **109**, 54.

Chambers, R. G., 1956, *Canad. J. Phys.*, **34**, 1395.

Chodorow, M., 1939, *Phys. Rev.*, **55**, 675.

Cohen, M. H., 1958, *Phil. Mag.*, **3**, 762.

COHEN, M. H., and HEINE, V., 1958, *Advanc. Phys.*, **7**, 395.
DANK, M., and CALLEN, H. B., 1952, *Phys. Rev.*, **86**, 622.
DONOVAN, B., and MARCH, N. H., 1956, *Proc. Phys. Soc.* B, **69**, 1249.
DRESSELHAUS, G., 1955, *Phys. Rev.*, **100**, 580.
DRESSELHAUS, G., KIP, A. F., and KITTEL, C., 1955, *Phys. Rev.*, **98**, 368.
ELLIOTT, R. J., 1954, *Phys. Rev.*, **96**, 266, 280.
ENGELMANN, F., 1956, *Z. Phys.*, **145**, 430.
FLETCHER, G. C., 1952, *Proc. Phys. Soc.* A, **65**, 192.
FLETCHER, G. C., and WOHLFARTH, E. P., 1951, *Phil. Mag.*, **42**, 106.
FLODMARK, S., 1959, *Ark. Fys.*, **14**, 513.
FUES, E., and STATZ, H., 1952, *Z. Naturforsch.*, **7a**, 2.
FUKUCHI, M., 1956, *Progr. Theor. Phys.*, **16**, 222.
GARCÍA MOLINER, F., 1958 a, *Proc. Phys. Soc.*, **72**, 996; 1958 b, *Phil. Mag.*, **3**, 207.
GLASSER, M. L., 1959, *J. Phys. Chem. Solids*, **10**, 229.
GLASSER, M. L., and CALLAWAY, J., 1958, *Phys. Rev.*, **109**, 1541.
GOLD, A. V., 1958, *Phil. Trans. Roy. Soc.* A, **251**, 85.
GOMBÁS, P., 1949, *Die statistische Theorie des Atoms und ihre Anwendungen* (Vienna: Springer).
GRIMLEY, T. B., 1958, *Proc. Phys. Soc.*, **71**, 749.
HALL, G. G., 1952, *Phil. Mag.*, **43**, 338; 1953, *Proc. Phys. Soc.* A, **66**, 1162; 1956, *Proc. Phys. Soc.* B, **69**, 1124; 1958, *Phil. Mag.*, **3**, 429; 1959, *Rep. Progr. Phys.*, **22**, 1 (London: Physical Society).
HAM, F. S., 1955, *Solid State Phys.*, **1**, 127.
HEINE, V., 1957, *Proc. Roy. Soc.* A, **240**, 340, 354, 361.
HERMAN, F., 1952, *Phys. Rev.*, **88**, 1210; 1954, *Ibid.*, **93**, 1214; 1955 a, *Proc. Inst. Radio Engrs*, **43**, 1703; 1955 b, *J. Electronics*, **1**, 103; 1958, *Rev. Mod. Phys.*, **30**, 102.
HERMAN, F., and CALLAWAY, J., 1953, *Phys. Rev.*, **89**, 518.
HERMAN, F., CALLAWAY, J., and ACTON, F. S., 1954, *Phys. Rev.*, **95**, 371.
HERRING, C., 1940, *Phys. Rev.*, **57**, 1169.
HERRING, C., and HILL, A. G., 1940, *Phys. Rev.*, **58**, 132.
HOWARTH, D. J., 1953, *Proc. Roy. Soc.* A, **220**, 513; 1955, *Phys. Rev.*, **99**, 469.
HOWARTH, D. J., and JONES, H., 1952, *Proc. Phys. Soc.* A, **65**, 355.
HOWLAND, L. P., 1958, *Phys. Rev.*, **109**, 1927.
HUBBARD, J., 1954, *Proc. Phys. Soc.* A, **67**, 1058.
JENKINS, D. P., 1954, *Physica*, **20**, 967; 1956, *Proc. Phys. Soc.* A, **69**, 548.
JENKINS, D. P., and PINCHERLE, L., 1954, *Phil. Mag.*, **45**, 93.
JONES, H., 1955, *Proc. Phys. Soc.* A, **68**, 1191.
JONES, H., and SCHIFF, B., 1954, *Proc. Phys. Soc.* A, **67**, 217.
KANE, E. O., 1956, *J. Phys. Chem. Solids*, **1**, 82; 1957, *Ibid.*, **1**, 249; 1959, *Ibid.*, **8**, 38.
KOBAYASI, S., 1958, *J. Phys. Soc. Japan*, **13**, 261.
KOHN, W., 1952, *Phys. Rev.*, **87**, 472; 1957, *Ibid.*, **105**, 509; 1958, *Ibid.*, **110**, 857; 1959, *J. Phys. Chem. Solids*, **8**, 45.
KOHN, W., and ROSTOKER, N., 1954, *Phys. Rev.*, **94**, 1111.
KORRINGA, J., 1947, *Physica*, **13**, 392.
KOSTER, G. F., 1953, *Phys. Rev.*, **89**, 67; 1957, *Solid State Phys.*, **5**, 174.
KUHN, T. S., and VAN VLECK, J. H., 1950, *Phys. Rev.*, **79**, 382.
LAX, B., 1958, *Rev. Mod. Phys.*, **30**, 122.
LEIGH, R. S., 1956, *Proc. Phys. Soc.* A, **69**, 388; 1958, *Proc. Phys. Soc.*, **71**, 33.
LÖWDIN, P. O., 1950, *J. Chem. Phys.*, **18**, 365; 1951, *Ibid.*, **19**, 1570, 1579; 1956, *Advanc. Phys.*, **5**, 1.
LUTTINGER, J. M., 1956, *Phys. Rev.*, **102**, 1030.
MCCLURE, J. W., 1957, *Phys. Rev.*, **108**, 612.
MACFARLANE, G. G., and ROBERTS, V., 1955, *Phys. Rev.*, **97**, 1714; **98**, 1865.
MARCH, N. H., 1957, *Advanc. Phys.*, **6**, 1.
MASE, S., 1958, *J. Phys. Soc. Japan*, **13**, 434.
MASLEN, V. W., 1956, *Proc. Phys. Soc.* A, **69**, 734.

Matyáš, Z., 1952, *Czech. J. Phys.*, **1**, 3.
Melvin, M. A., 1956, *Rev. Mod. Phys.*, **28**, 18.
Miasek, M., 1956, *Bull. Acad. Polon. Sci. Cl.* 3, **4**, 453.
Morita, A., 1958, *Progr. Theor. Phys.*, **19**, 534 ; 1959, *J. Phys. Chem. Solids*, **8**, 363.
Morita, A., and Takahashi, K., 1958, *Progr. Theor. Phys.*, **19**, 257.
Mott, N. F., 1956, *Nature, Lond.*, **178**, 1205.
Nutkins, M. A. E., 1956, *Proc. Phys. Soc.* B, **69**, 619.
Parmenter, R. H., 1952, *Phys. Rev.*, **86**, 552.
Parratt, L. G., 1959, *Rev. Mod. Phys.*, **31**, 616.
Parzen, L. G., 1953, *Phys. Rev.*, **89**, 237.
Phillips, J. C., 1958, *Phys. Rev.*, **112**, 685 ; 1959, *J. Phys. Chem. Solids*, **8**, 369.
Phillips, J. C., and Kleinman, L., 1959, *Phys. Rev.*, **116**, 287.
Pincherle, L., 1958, *Proc. Phys. Soc.*, **72**, 281.
Pines, D., 1955, *Solid State Phys.*, **1**, 368.
Pippard, A. B., 1957, *Phil. Trans. Roy. Soc.* A, **250**, 325.
Raimes, S., 1954 a, *Phil. Mag.*, **45**, 727 ; 1954 b, *Proc. Phys. Soc.* A, **67**, 52.
Raychaudhuri, A., 1957, *Z. Phys.*, **148**, 435.
Raynor, G. V., 1952, *Rep. Progr. Phys.*, **15**, 173 (London : Physical Society).
Reitz, J. R., 1955, *Solid State Phys.*, **1**, 1 ; 1957, *Phys. Rev.*, **105**, 1233.
Roth, L. M., Lax, B., and Zwerdling, S., 1959, *Phys. Rev.*, **114**, 90.
Saffren, M. M., and Slater, J. C., 1953, *Phys. Rev.*, **92**, 1126.
Saint-James, D., 1958, *C. R. Acad. Sci., Paris*, **246**, 1533.
Sasaki, W., Shibuya, M., Mizuguchi, K., and Hatoyama, G. M., 1959, *J. Phys. Chem. Solids*, **8**, 250.
Saxon, D. S., and Hunter, R. A., 1949, *Philips Res. Rep.*, **4**, 81.
Schiff, B., 1954, *Proc. Phys. Soc.* A, **67**, 2 ; 1955, *Ibid.*, **68**, 686 ; **69**, 185.
Segall, B., 1957, *Phys. Rev.*, **105**, 108 ; 1959, *J. Phys. Chem. Solids*, **8**, 371.
Seitz, F., 1935, *Phys. Rev.*, **47**, 400.
Shakin, C., and Birman, J. L., 1958, *Phys. Rev.*, **109**, 818.
Shibuya, M., 1955, *Phys. Rev.*, **99**, 1189.
Shockley, W., 1937, *Phys. Rev.*, **52**, 866 ; 1950, *Ibid.*, **78**, 173.
Silverman, R. A., 1952, *Phys. Rev.*, **85**, 227.
Skinner, H. W. B., 1939, *Rep. Progr. Phys.*, **5**, 257 (London : Physical Society).
Slater, J. C., 1937, *Phys. Rev.*, **51**, 846 ; 1951 a, *J. Chem. Phys.*, **19**, 220 ; 1951 b, *Phys. Rev.*, **81**, 385 ; 1952, *Ibid.*, **87**, 807, 1952 ; 1953, *Ibid.*, **92**, 603 ; 1956, *Handbuch der Physik*, Vol. 19 (Berlin : Springer) ; 1959, *J. Phys. Chem. Solids*, **8**, 21.
Slater, J. C., and Koster, G. F., 1954, *Phys. Rev.*, **94**, 1498.
Slonczewski, J. C., and Weiss, P. R., 1958, *Phys. Rev.*, **109**, 272.
Soule, D. E., and McClure, J. W., 1959, *J. Phys. Chem. Solids*, **8**, 29.
Steinmann, O., 1957, *Helv. Phys. Acta*, **30**, 515.
Suffczynski, M., 1958, *Bull. Acad. Polon. Sci.* (*Phys., Math., Astr.*), **6**, 481.
Ten Seldam, C. A., 1957, *Proc. Phys. Soc.* A, **70**, 97, 529.
Trefftz, E., 1954, *Z. angew. Math. Mech.*, **34**, 303.
Trlifaj, M., 1952, *Czech. J. Phys.*, **1**, 110.
Von der Lage, F. C., and Bethe, H. A., 1947, *Phys. Rev.*, **71**, 612.
Wainwright, T., and Parzen, G., 1953, *Phys. Rev.*, **92**, 1129.
Wannier, G. H., 1937, *Phys. Rev.*, **52**, 191 ; 1938, *Ibid.*, **53**, 671.
Wigner, E., and Seitz, F., 1933, *Phys. Rev.*, **43**, 804 ; 1934, *Ibid.*, **46**, 509 ; 1955, *Solid State Phys.*, **1**, 97.
Winston, H., 1954, *Phys. Rev.*, **94**, 328.
Woodruff, T. O., 1955, *Phys. Rev.*, **98**, 1741 ; 1956, *Ibid.*, **103**, 1159 ; 1957, *Solid State Phys.*, **4**, 367.

ORBITAL ELECTRON CAPTURE BY THE NUCLEUS

By R. BOUCHEZ and P. DEPOMMIER
Nuclear Physics Laboratory, University of Grenoble, France

CONTENTS

PAGE

§ 1. General properties of electron capture 396
1.1. Historical account 397
1.2. Energy considerations 398
1.2.1. Conservation of energy 398
1.2.2. Orbital capture and β^+ emission 400
1.2.3. Determination of the energy available from capture 401
1.3. Characteristic radiations 404
1.3.1. X-rays from orbital capture 405
1.3.2. Auger electrons from orbital capture 406
1.3.3. Measurements of x-rays and Auger electrons 410
§ 2. The neutrino 411
2.1. 'Old' properties of the neutrino 412
2.2. 'New' properties of the neutrino 413
2.3. Asymmetry of the neutrino 414
2.3.1. Dirac theory with four components 414
2.3.2. Pauli two-component theory 415
2.3.3. Parity, chirality and helicity 416
2.3.4. Four or two neutrinos 416
2.3.5. Lepton conservation 417
2.3.6. The two-component theory 417
2.3.7. The new Majorana hypothesis 418
§ 3. Fermi interaction and orbital capture 418
§ 4. Transition probabilities for capture 421
4.1. General expression of the transition probabilities 421
4.2. Selection rules 422
4.3. Nuclear matrix elements and $f_K T_K$ products 425
4.3.1. Form of the matrix elements 425
4.3.2. $f_K T_K$ products 426
4.3.3. Classification of the transitions according to the $f_K T_K$ product 427
4.4. Relativistic eigenfunctions of the leptons 427
4.4.1. Atomic electron radial functions 427
4.4.2. Neutrino radial functions 428
4.4.3. Positron radial functions 430
§ 5. The allowed transitions 430
5.1. Møller's formulae 431
5.1.1. Allowed transitions $\Delta I = 0$ (not $0 \rightarrow 0$) 431
5.1.2. Allowed transitions $0^{\pm} \rightarrow 0^{\pm}$ 432
5.1.3. Allowed transitions $\Delta I = 1$ 432
5.2. Experimental results 432
§ 6. Forbidden transitions with 'unique' matrix element 434
6.1. Theoretical formulae 434
6.2. Typical experimental transition: ^{125}Sb (2·73d) 435
§ 7. Parity-forbidden transitions 436
7.1. Parity-forbidden transitions $\Delta I = 1$ 436
7.2. Parity-forbidden transitions $\Delta I = 0$ 437
7.3. Typical experimental transition: ^{238}Np 438
§ 8. Forbidden transitions of nth order 438

PAGE

§ 9. L capture and K capture 438
9.1. Theoretical results for L/K ratio 438
9.1.1. Allowed transitions 440
9.1.2. Forbidden 'unique' transitions $\Delta I = n+1$, $\Delta\pi = (-1)^n$ 440
9.1.3. Parity-forbidden transitions 440
9.2. Experimental results for L/K ratio—comparison with theory 440
9.2.1. Allowed transitions 440
9.2.2. Parity-forbidden transitions 442
9.2.3. Conclusion 443
§ 10. Capture positron ratio 444
10.1. K/β^+ ratio for allowed transitions 444
10.1.1. Theoretical results 444
10.1.2. Experimental results—comparison with theory 445
10.2. K/β^+ ratio for forbidden transitions 445
10.2.1. Forbidden 'unique' transitions $\Delta I = n+1$, $\Delta\pi = (-1)^n$ 445
10.2.2. K/β^+ ratio for first-forbidden transitions 447
10.3. Conclusion 447
References 449

Abstract. The general properties of orbital electron capture are reviewed. Energetic considerations, radiations associated with capture (x-rays and Auger electrons) and their experimental investigation are discussed.

Taking into account the new results on the β interaction, formulae for transition probabilities for any order of forbiddenness have been calculated by means of the spherical tensor method and using the two-component neutrino theory with (V, A) interaction.

L/K ratios have been computed from these formulae, using the bound electron wave functions given by Brysk and Rose. These theoretical results have been compared with the experimental data and the agreement is fairly good.

K/β^+ ratios have also been computed, using Dzelepov's tables for the β^+ spectrum and Brysk's and Rose's functions for the K electron. For allowed and 'unique' transitions the calculated values agree with experiment.

§ 1. General Properties of Electron Capture

The simple fact of the capture of an atomic electron by a proton inside the nucleus points to the influence of the atomic electrons on nuclear properties, and further brings direct evidence for the interaction between proton and electron. A study of orbital capture must approach the event in its nuclear origin and in its atomic influences.

The nuclear part of the event is similar to β emission in the sense that orbital capture is essentially a neutron–proton transition initiated by the coupling between nucleons and leptons.† Then capture allows a good analogy with β emission.

The atomic influences on capture are twofold. The probability of electron capture depends directly on the characteristics of the captured electron, which are influenced by the size of the nucleus, by the distribution of the protons inside the nucleus and of the other electrons of its own atom, or even of the neighbouring atoms. Further, electron capture is followed by a rearrangement of the atomic shells, producing x-rays and Auger electron emission, whose analysis is the most straightforward method for the investigation of capture.

† According to Rosenfeld's terminology, the light particles, electron and neutrino, are called leptons.

Orbital capture, like β emission, depends on the parameters (energy E, ' spin ' I,† parity π) characterizing the initial and final nucleus. It depends also on the modification of nuclear structure in the transition.

This nuclear structure effect is characterized precisely by the nuclear matrix element of the transition, or approximately by the value of the fT product which can be obtained from the Sargent curves (1933) (Major and Biedenharn 1954), or by direct calculation.

Orbital capture, therefore, cannot be separated from β emission in the investigation of nuclear structure. But it does give information on the eigenfunctions of the atomic electrons and on the rearrangement of the electronic shells. The experimental parameters which serve as basis for every precise analysis of capture are either the intensity of the K and L capture transitions (L/K, L_{II}/L_I or M/L ratio), or the energy and the intensity of x-radiation and Auger electrons.

Finally, in the case of β^+ emission, the K/β^+ branching ratio is an excellent test of β disintegration theory.

1.1. *Historical Account*

Electron capture was predicted theoretically as early as 1935, before being observed in 1937. It is however interesting to note that Irene Curie and F. Joliot had as early as 1933 anticipated electron capture radioactivity: in their study (I. Curie and Joliot 1933) of neutron production in (α, n) reactions they supposed that the unknown isotope ^{22}Na produced in the ^{19}F (α, n) reaction might transform spontaneously into ^{22}Ne by capture of an atomic electron. We now know that ^{22}Na decays by orbital capture with a branching ratio of about 10%.

The theoretical investigation of capture was carried out independently by Yukawa and Sakata in 1935, Bethe and Bacher in 1936 and Møller in 1937 (a, b); these authors applied to capture the theory of β emission which Fermi (1934) had just elaborated.

The evidence for K capture was established in 1937 by Alvarez, for ^{48}V (16 days); he detected x-rays TiK_α emitted during the rearrangement of the K shell of titanium. The idea of detecting characteristic x-rays from the capture was first suggested by Jacobsen (1937) and applied without success to ^{45}Sc; the ^{45}Sc source produced by the Copenhagen cyclotron was too weak and Jacobsen could not observe the photoelectrons produced by x-rays in a Wilson cloud chamber. On the contrary, Alvarez operating the more powerful cyclotron in Berkeley made obvious the TiK photons by means of a thin-walled Geiger–Müller counter filled with argon at atmospheric pressure.

Since the discovery of the K capture of ^{48}V about 150 isotopes decaying by capture have been found. The most frequently observed capture is that of a K electron.

But capture of an L electron can also occur; sometimes it is even more probable than K capture, as with ^{40}K (Heintze 1954). L capture may also occur alone, as seems to be the case with ^{185}Os (Miller and Wilkinson 1951). Ignored for a long time, the theory of L and further captures was worked out by Marshak (1942), but L capture was discovered only in 1949 by Pontecorvo *et al.* for ^{37}A by means of a proportional counter. Let us point out, however, that the evidence

† The quantum number of total angular momentum **I** is often called ' spin '. $\mathbf{I}^2\phi = I(I+1)\hbar^2\phi$.

for a difference between the periods of ^{7}Be metal and its fluoride (Leininger, Segrè and Wiegand 1949, Bouchez *et al.* 1949, Bouchez 1950) can be considered as the first indirect observation of L capture : in this experiment, it can be said in first approximation that in the metal the overall capture probability is $P_K + P_L$ whereas in the fluoride the overall capture probability of $^7Be^{2+}$ ion is only P_K, whence a variation of the radioactive period $\Delta T/T \simeq P_L/P_K$ due to L capture is to be expected.

1.2. *Energy Considerations*

1.2.1. *Conservation of energy.*

Electron capture is essentially the capture of an atomic electron by a proton inside the nucleus, as a result of which the proton is transformed into a neutron. The greater part of the energy released in this transition is not detected, and by analogy with Pauli's hypothesis about β emission, it is postulated that a neutrino carries off the energy difference ; a transition by capture can be written

$$(p) + e^- \rightarrow (n) + \nu \qquad \ldots\ldots(1)$$

(p) and (n) indicating that proton and neutron are bound to the other nucleons in the nucleus, and ν being the emitted neutrino following orbital capture.

The energy balance of the transition by capture corresponding to (1) can be written

$$M_i + \epsilon_n = M_f + q_n + \epsilon_R \qquad \ldots\ldots(2)$$

M_i and M_f being the ' nuclear ' masses of the initial and final nucleus, ϵ_n the whole energy of the captured electron in the nth state (one has approximately $\epsilon_n = 1 - |E_n|$, $|E_n| = \alpha^2 Z_e^{\,2}/n^2$ being the binding energy of the captured electron in units $m_0 c^2$, Z_e the effective atomic number taking screening into account and m_0 the rest mass of the electron) ; q_n is the whole energy carried by the neutrino in the final state and ϵ_R the recoil kinetic energy of the final nucleus.

Relation (2) can also be written, if the difference of nuclear masses $(M_i - M_f)$ is put equal to ΔM_n

$$\Delta M_n + \epsilon_n = q_n + \epsilon_R. \qquad \ldots\ldots(3)$$

This energy balance (3) draws attention to the magnitude $\Delta M_n + \epsilon_n$ representing the energy available from the transition and carried off by the neutrino and the recoil nucleus.

The rest mass of the neutrino is very small. According to the experiments of Langer and Moffat (1952) and Hamilton, Alford and Gross (1953) on the shape of the β spectrum of ^{3}H close to the maximum energy, it is not greater than $10^{-3}\, m_0$. Also, recent experiments on the non-conservation of parity in β radioactivity indicate that the neutrino mass should be entirely kinetic (in analogy with the photon except for the difference in spin). It is therefore assumed that the rest mass of the neutrino is zero.

The recoil kinetic energy is always very small. It can be obtained from the equation of momentum conservation, and is equal to

$$\epsilon_R = \frac{q^2}{2Mc^2} \simeq 537\,\frac{q^2}{A}\,\text{ev},$$

ϵ_R being given in ev and q in Mev (this relation holds for every particle with zero rest mass, i.e. photon, neutrino). This recoil energy is large for light nuclei

(low A) and for appreciable neutrino energies. The greatest value (57 ev) is obtained from the transition $^7\text{Be} \xrightarrow{K} {}^7\text{Li}$.

If the recoil energy is neglected, the energy available from capture is in practice carried off by the neutrino (if there is no γ emission) and takes the simple form

$$\Delta M_n + \epsilon_n \simeq q_n. \qquad \ldots\ldots(4)$$

On the contrary, it is known that in β emission the energy is shared between electron and neutrino, each particle having therefore a continuous energy spectrum.

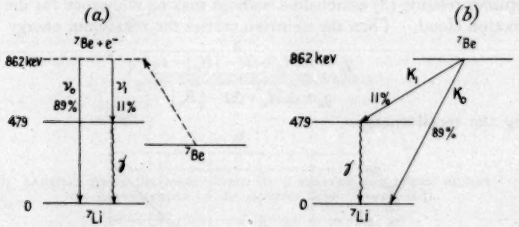


Figure 1. ^{7}Be decay scheme by capture with (*a*) nuclear masses or (*b*) atomic masses.

As an example, let us apply the energy balance to the case of the transition $^7\text{Be} \rightarrow {}^7\text{Li}$ (figure 1). The difference ΔM_n of the nuclear masses can be obtained from the threshold of the $^7\text{Li}(p, n)^7\text{Be}$ reaction with $Q = -1{\cdot}645$ Mev. One gets

$$\Delta M_n = {}^7\text{Be} - {}^7\text{Li} = Q + p - n = 351 \text{ kev}.$$

The maximum energy carried by the neutrino in this capture is therefore $\Delta M_n + \epsilon_K = 862$ kev, neglecting the binding energy of the K electron, which is of the order of magnitude of 100 ev ($\epsilon_K \simeq m_0 c^2 = 511$ kev).

The above energy balance has been written without taking into account the electron cloud. However, the $Z-1$ electrons in the final atomic state do not have the same coulomb potential as in the initial state ; there arises an expansion of the electron cloud following the charge variation $Z \rightarrow Z-1$: the final atomic states not being identical with the initial atomic states. The detailed study of the role of the electron cloud in capture has been worked out by Benoist-Gueutal (1953) : a first correction to the transition probability occurs especially for light atoms, like ^{7}Be, for which the relative variation from Z to $Z-1$ is largest. As a result the transition probability is reduced by about 1% for ^{7}Be, the effect being negligible for heavier atoms like ^{18}F, ^{22}Na, ^{37}A. A second correction is necessary on account of the change in atomic binding energy ; the charge variation accompanying capture also causes a reduction of the atomic binding energy which can be computed for rather heavy atoms according to the statistical model of Thomas–Hartree, giving for the mean atomic binding energy the value $\bar{\epsilon} \simeq -RZ^m$ in units $m_0 c^2$, where $R = \alpha^2/2 \simeq 13{\cdot}6$ ev is Rydberg's constant and m a parameter about 2·4. This reduction of binding energy corresponds to the absorption of energy $\Delta\bar{\epsilon} \simeq mRZ^{m-1}$ of some kev (figure 2), taken from the total energy of transition.

Taking into account this rearrangement of the electron shells, the energy balance applied to the atom between its initial state and its final state in their ground states gives

$$M_a^Z = M_a^{Z-1} + |E_n| + q_n + \epsilon_R - \Delta\bar{\epsilon} \qquad \ldots\ldots(5)$$

M_a^Z and M_a^{Z-1} being the atomic mass of the initial and the final atom in their ground state. The relation (5) can also be written

$$\Delta M_a - |E_n| = q_n + \epsilon_R - \Delta\bar{\epsilon} \qquad \ldots\ldots(6)$$

which replaces relation (3) established without making allowance for the expansion of the electron cloud. Then the neutrino carries the maximum energy

$$q_n = \Delta M_a + \Delta\bar{\epsilon} - |E_n| - \epsilon_R \qquad \ldots\ldots(7a)$$

or

$$q_n \simeq \Delta M_a + \Delta\bar{\epsilon} - |E_n| \qquad \ldots\ldots(7b)$$

neglecting the recoil energy.

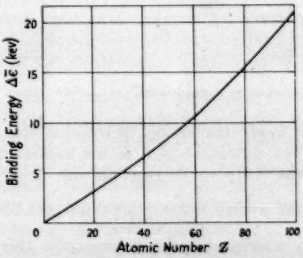


Figure 2. Decrease of binding energy of the atomic electrons in a transition by capture or β^+ emission. (*After* Benoist-Gueutal 1953.)

1.2.2. *Orbital capture and β^+ emission.*

To permit orbital capture by the nucleus, the mass of the initial atom must from (7) be at least equal to the mass of the final atom; whence the condition

$$\Delta M_a \geqslant |E_n| - \Delta\bar{\epsilon} \quad \text{or} \quad \Delta M_a \gtrsim 0. \qquad \ldots\ldots(8)$$

When $0 \lesssim \Delta M_a \lesssim 2m_0c^2$, β^+ emission does not occur. But, when the energy condition $\Delta M_a \gtrsim 2m_0c^2$ for β^+ decay is fulfilled, it is *a fortiori* for capture (figure 3). So, orbital capture always accompanies positron emission. For each β^+ emitter an important experimental parameter is the K/β^+ ratio of the relative probabilities of decay by capture and β^+ emission.

The region of stability between two neighbouring isobars is very narrow, if it exists, and is always less than the binding energy $|E_n|$ of the least bound atomic electron. In practice there exist in the chart of elements, at most two pairs of stable isobars :

^{113}Cd(12·3%)–In(4·2%) and ^{123}Sb(42·7%)–Te(0·87%).

But, according to the results of Cohen (1951) ^{113}In should decay by L capture with a very long lifetime; however according to the analysis of energy parabolae (Bouchez *et al.* 1953 a) ^{113}Cd might be a long-lived β^- emitter. Let us note that this stability region between β^- emission and K capture could also be reduced by e-creation (Daudel *et al.* 1947), that is to say β^- emission in which the emitted electron remains bound to the atom and is created with an energy slightly less than

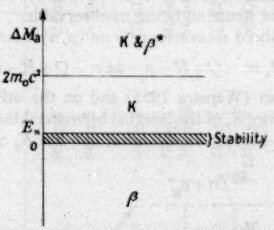


Figure 3. Diagram giving the energy limits for β emission and orbital capture (the rearrangement of the electron cloud is neglected).

$m_0 c^2$. It seems then probable that no pair of stable isobars can exist but that one member of the pair is radioactive (β^- or capture) with a very long lifetime.

It should be pointed out that if one considers two stable isobars, A even, with their atomic numbers Z and $Z+2$ differing by two units, the intermediate isobar transforms (figure 4) both by β^- emission and by capture (Sizoo 1937) or β^+ emission if enough energy is available.

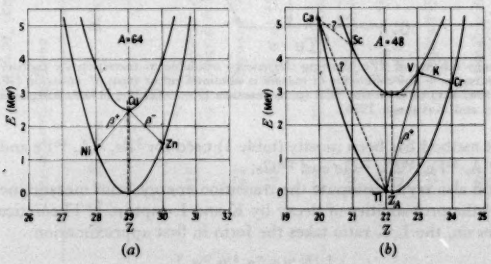


Figure 4. (*a*) Decay scheme of ^{64}Cu (β^- and β^+ emitter). Interpretation by the energy parabolae (Bouchez *et al.* 1953 a). (*b*) $A = 48$ isobars. Analogy with ^{64}Cu, but the shift of the maximum of stability ($Z_A \simeq 22{\cdot}2$) towards the stable isobar ^{48}Ti makes the isobar ^{48}Cr unstable.

1.2.3. *Determination of the energy available from capture.*

When a nucleus can decay by positron emission as well as by capture, the maximum kinetic energy W_0 of the emitted positrons makes it possible to obtain the energy available from the capture (figure 5) which leads to the same level as

β^+ transition. Indeed, the conservation of energy between the atomic states gives for β^+ emission

$$\Delta M_a = W_0 + 2m_0c^2 + \epsilon_R - \Delta\bar{\epsilon} \qquad \ldots\ldots(9)$$

analogous to relation (6) for capture.

Let us remark that the detection of positrons or of annihilation photons provides us with a lower limit of the energy q_n carried by the neutrino.

When β^+ emission cannot occur, it is more difficult to determine the energy available from capture, the neutrino being unobservable. The principal methods are on the one hand threshold measurements on (p, n) reactions

$$\Delta M_a = -Q + H - n - \Delta\bar{\epsilon} \simeq -Q + H - n \qquad \ldots\ldots(10)$$

where $n - H = (783 \pm 1)$ kev (Wapstra 1955) and on the other hand, the measurement of the maximum energy E_γ of the internal bremsstrahlung photons accompanying capture (Michalowicz 1956). In this case, one has $E_\gamma \simeq q_n$.

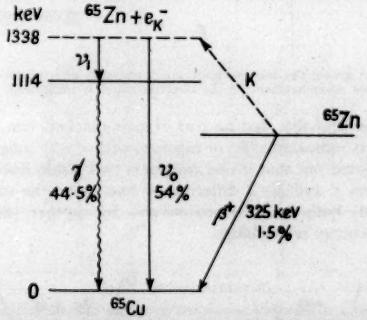


Figure 5. Decay scheme of ^{65}Zn showing the energy available in transition by capture in the case of β^+ emission. The difference of masses is obtained either from β^+ emission ($W_0 = 325$ kev, $\Delta M_a \simeq 1347$ kev) or from the ^{65}Cu (p, n) reaction $Q = -(2131 \pm 1)$ kev, $\Delta M_a = (1349 \pm 2)$ kev (Marion and Kavanagh 1956).

The first method has been mostly (table 1) used for ^{7}Be, ^{37}A, ^{55}Fe and ^{65}Zn, the second for ^{37}A, ^{55}Fe, ^{59}Ni, ^{71}Ge and ^{131}Cs.

One could also try to compute the transition energy from measurements of the L/K ratio of the probabilities of decay by K and L capture. Theoretically, as we shall see later on, the L/K ratio takes the form in first approximation

$$L/K \simeq q_L^2 g_L^2 / q_K^2 g_K^2 \qquad \ldots\ldots(11)$$

$q_L \simeq \Delta M_a - |E_L|$ and $q_K \simeq \Delta M_a - |E_K|$ being the energy carried by the neutrino in K and L capture respectively; and g_L^2 and g_K^2 being the electron densities at the nuclear surface. But very often $q_L \simeq q_K$ and as a result the L/K ratio is not very sensitive to a variation of ΔM_a and differs only very little from the g_L^2/g_K^2 ratio. Moreover the determination of L/K sometimes makes use of the fluorescence yield R_K which is not known to great accuracy. Finally, to get L/K experimentally it is necessary to analyse low energy x-rays and Auger electrons. Measurements

Table 1. Energy of the Transition by Capture characterized by the Difference ΔM_a (kev) of the Atomic Masses

Nuclide	Internal bremsstrahlung			(p, n) threshold		
	E_γ (kev)	$\Delta M_a = E_\gamma + E_K$	Reference	$-Q$ (kev)	$\Delta M_a - Q - 783$	Reference
^{7}Be				1645	862 ± 2	Richards *et al.* 1950
^{37}A	815 ± 20	818	Emmerich *et al.* 1954 Linqvist & Wu 1955	1596	813 ± 2	Schoenfeld *et al.* 1952
^{48}V	617 ± 10	622	Hayward 1955	1391	608 ± 5	Trail & Johnson 1953
^{51}Cr	751 ± 5	757	Bisi *et al.* 1955	1532 1540	752 ± 6 760 ± 10	Richards *et al.* 1950 Stelson *et al.* 1950
^{55}Fe	215 ± 10	222	Michalowicz 1956	1020	237 ± 5	Trail & Johnson 1953
^{66}Ni	1065 ± 30	1073	Emmerich *et al.* 1954	1857	1074 ± 3	McCue & Preston 1951
^{71}Ge	226 ± 5 220 ± 3	237 231	Langevin 1955 a Bisi 1955	1030	247 ± 30	Trail & Johnson 1953
^{103}Pd	494^{+27}_{-12}	518	Rietjens 1956			
^{131}Cs	325 ± 15 320 ± 10	361 356	Michalowicz 1956 Saraf 1954			

are difficult and the accuracy is rather poor. It is then difficult to make use of the L/K ratio to compute the energy of transition by capture.

An order of magnitude for ΔM_a can also be obtained from the energy parabolae, especially for nuclei with odd A the parabolae of which are well enough known.

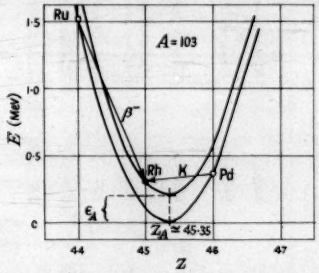


Figure 6. Energy of the transition by capture $^{103}Pd \rightarrow {}^{103}Rh$, obtained from the energy parabolae. The fine structure of the parabolae ($\epsilon_A \simeq 200$ kev) is due to the fact that a proton pair is generally more tightly bound than a neutron pair.

This method applied to the case of ^{103}Pd (Bouchez *et al.* 1955) made it possible to get $\Delta M_a \simeq 600$ kev, in good agreement with the analysis of γ-rays by Rietjens (1956) who obtains 518 kev $\pm$ 25 kev.

1.3. *Characteristic Radiations*

Orbital capture manifests itself only in the form of x-rays and Auger electrons which are emitted during the rearrangement of the electron shells. Very large intensities of x-rays and soft electrons are characteristic of transitions by capture.

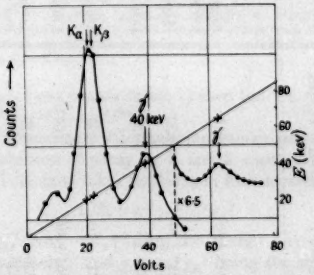


Figure 7. Spectrum of the x-rays and soft γ-rays from the capture by ^{103}Pd (17d) given by a scintillation counter. A foil of Cu ($\frac{1}{15}$ mm) absorbs a part of the Rh K x-radiation ($\simeq 20$ kev). (*After* Avignon *et al.* 1955.)

We shall not speak of the internal bremsstrahlung associated with every transition by capture. Figure 7 shows the radiation from ^{103}Pd which was obtained by

means of a scintillation counter. Let us note the large intensity of Rh K_α (20·3 kev) and Rh K_β (22·7 kev) x-radiations corresponding to about 0·8 photon per disintegration (Avignon *et al.* 1955, de Raad *et al.* 1954), while ^{103}Pd emits only about $1{\cdot}7 \times 10^{-5}$ γ-ray (40 kev) per disintegration. In practice a ^{103}Pd source of 10 mc emits 8 mc of x-rays between 20 and 22 kev, which allows the use of this isotope as a portable x-ray source.

The large intensity of Auger electrons associated with capture is illustrated by Cook and Langer's experiments (1948–1951) on ^{64}Cu (figure 8).

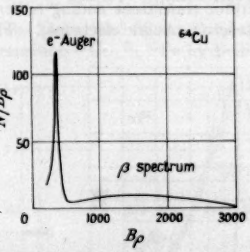


Figure 8. Spectrum of the electrons emitted by ^{64}Cu. The continuous distribution corresponds to the normal β spectrum of ^{64}Cu and the weak energy peak is due to the Auger electrons connected with capture. (*After* Cook 1951.)

1.3.1. *X-rays from orbital capture.*

The final atom, ionized in the K level by K capture, can emit the x-ray lines of the K series† chiefly $K_{\alpha_1}(KL_{III})$, $K_{\alpha_2}(KL_{II})$, $K_{\beta_1}(KM_{III})$ and $K_{\beta_2}(KM_{II})$ with relative intensities (Compton and Allison 1935)

$$\alpha_1 : \alpha_2 : \beta_1 : \beta_2 \simeq 100 : 48 : 18 : 9.$$

After the emission of K_{α_1} and K_{α_2} lines, the atom is ionized in the L_{III} or L_{II} levels and can emit x-rays of the L series $L_{\alpha_1}(L_{III}\,M_V)$, $L_{\alpha_2}(L_{III}\,M_{IV})$, $L_{\beta_1}(L_{II}\,M_{IV})$. These L radiations correspond to one L_{II} and L_{III} initial ionization, but in practice not to an L_I ionization, the K–L_I transition being forbidden. The result is that after K capture only very few atoms are finally ionized in the L_I level.

On the contrary, an L capture leaves the final atom mostly ionized in the $L_I(2s)$ level, p capture ($L_{II}\,L_{III}$) being very improbable (except for a few forbidden transitions); the atom can then emit x-rays leading to the L_I level: $L_{\beta_2}(L_I\,M_{III})$, $L_{\beta_3}(L_I\,M_{II})$. Let us note that L_I ionization is therefore in first approximation characteristic of L capture, and L_{II} or L_{III} ionization characteristic of K capture. However, the investigation of L capture by this method is difficult; on the one hand the analysis of L x-rays is not easy and on the other hand, $L_I\,L_{II}$ or $L_I\,L_{III}$ transitions by the Coster–Kronig (1935) effect can leave the atom ionized in the L_{II} or in the L_{III} level after the capture of an L_I electron.

† The energy of the x-rays can be computed from tables by Cauchois and Hulubei (1947) or Hill, Church and Mihelich (1952).

1.3.2. *Auger electrons from orbital capture.*

An atom ionized in the K level, after K capture or internal conversion, goes back to its ground state by emission of either an x-ray belonging to the K series (K_{α_1}, K_{α_2}), or an electron from an outer shell, which carries off the available energy : it is the Auger effect (1925, 1926), characterized by the fluorescence yield R_K = (number of K photons emitted)/(number of K ionizations). The Auger effect corresponds accordingly to an electronic transition without emission of radiation and with energy transfer to an electron in an outer level.

The Auger electrons accompany not only K capture but also L capture and are associated with the electronic transitions ending either on the K level (K Auger electrons) or on the L level (L Auger electrons). The reader will find a good

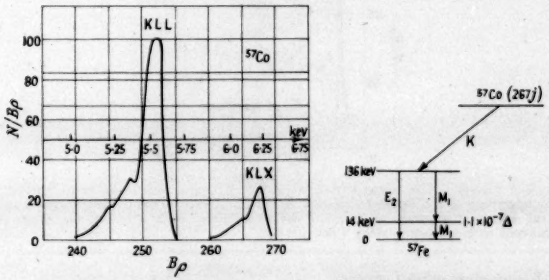


Figure 9. Auger electron groups KLL and KLX associated with K capture and internal conversion of ^{57}Co. (*After* Bellicard and Moussa 1957.)

analysis of the Auger effect and Auger electrons in Burhop's book (1952) and in Bergström's papers (Bergström and Hill 1954, Bergström 1955). We shall sum up here the main characteristics of these effects, in relation to orbital capture.

Energy of the Auger electrons. The K Auger electrons fall principally into three energy groups :

(*a*) The most intense is the KLL group, corresponding to the L electrons emitted after a K–L transition ; it is the group of lowest energy.

(*b*) The intermediate group KLX made up of the X electrons (coming from a level higher than L) emitted after a K–L transition.

(*c*) The KXY group, of highest energy, and also of lowest intensity, which is emitted after a K–X transition and corresponds to Y electrons belonging to an X shell or an outer shell. Figure 9 shows the Auger electrons of the main groups KLL and KLX emitted in the disintegration of ^{57}Co after K capture or K internal conversion (Bellicard and Moussa 1957).

If the K Auger electrons have been studied very frequently, the L Auger electrons, on the contrary, have been investigated much less, their energy being too low. The L Auger electrons associated with an ionization of the L_I, L_{II} or L_{III} levels can be separated (less distinctly than the K Auger electrons) into LMX, LMM and LXY electrons.

A further particular group due to the Coster–Kronig effect corresponds for instance to the M_{IV} or M_V electrons which are emitted after the $L_I L_{III}$ transition. These Coster–Kronig electrons can be emitted only for $Z \leqslant 50$ and $Z \geqslant 80$ (Burhop 1952), the binding energy of the M electrons being lower than the difference L_I–L_{III} (figure 10).

The kinetic energy of the Auger electrons can be computed to a first approximation from the binding energies K(Z) of the levels. For instance, the energy of the $KL_{II}\,L_{III}$ Auger electrons is given by

$$E_a(KL_{II}\,L_{III}) \simeq K(Z) - L_{II}(Z) - L_{III}(Z). \qquad \ldots\ldots(12a)$$

In the case of iron the relation (12*a*) gives 5·68 kev ; and the experimental measurement, carried out on the transition $^{57}Co \xrightarrow{K} {}^{57}Fe$ by Bellicard and Moussa (1957),

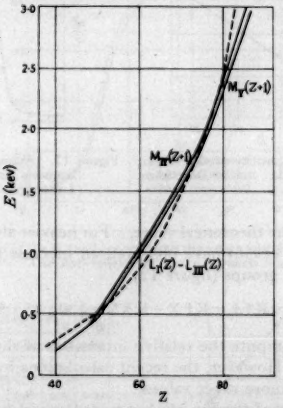


Figure 10. Comparison between the binding energy of the M_{IV} and M_V electrons and the difference L_I–L_{III} indicating the regions where Coster–Kronig transitions are possible. (*After* Burhop 1952.)

gives the value (5·57 ± 0·03) kev, lower than the theoretical approximate value. A more exact calculation must take into account the double ionization of the atom after the Auger electron left it ; relation (12*a*) must be replaced by (Burhop 1952, Bergström 1955)

$$E_a(KL_{II}\,L_{III}) = K(Z) - L_{II}(Z) - L_{III}(Z + \Delta Z) \qquad \ldots\ldots(12b)$$

where $0 \leqslant \Delta Z \leqslant 1$. Applying (12*b*) with $\Delta Z = 1$ to the Auger electrons of iron, the value 5·61 kev is found, which is in better agreement with the experimental value.

Intensity of the Auger electrons. The relative intensities of the Auger groups KLL, KLX and KXY are still little known. The probabilities have been calculated by Pincherle (1935) in the non-relativistic case and with hydrogen-like functions,

without taking screening into account. These probabilities therefore do not depend on Z and have the approximate values

$$\text{KLL} : \text{KLM} : \text{KMM} \simeq 100 : 70 : 10.$$

However, the experimental values (Bergström 1955) (figure 11) vary with atomic number and differ considerably from Pincherle's values ; in the case of the transition by capture of ^{65}Zn, the ratio KLL/KLX $\simeq$ 100/22 (accurate to 5%) (Bouchez *et al.* 1953) is in agreement with the experimental values obtained by Bellicard and Moussa (1957) for the transition $^{51}\text{Co} \xrightarrow{K} {}^{57}\text{Fe}$: KLL/KLX $\simeq$ 100/22 (accurate to 5%) and

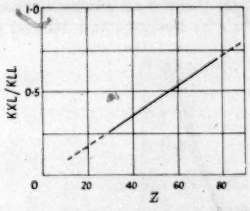


Figure 11. Relative intensities of the Auger groups KLX/KLL. (*After* Bergström 1955.)

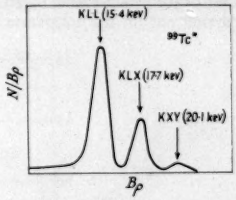


Figure 12. Auger groups K obtained in the isomeric transition of ^{99}Tc. (*After* Laberrigue-Frolow and Radvanyi 1956.)

is very different from the theoretical value. For heavier atoms ($Z \simeq 45$) Laberrigue-Frolow and Radvanyi (1956) have observed an increase of the relative intensities of the KLX and KXY groups (figure 12) :

$$\text{KLL} : \text{KLX} : \text{KXY} \simeq 100 : 45 : 5.$$

Thus it is better to compute the relative intensities of the Auger groups from the experimental results. However, the recent calculations by Rubenstein and Snyder (1955) should lead to more exact values.

The whole intensity of the K or L Auger electrons depends essentially on the K and L fluorescence yields R_K and R_L. In certain experiments on the branching ratios K/β^+ and L/K it is necessary to know exactly the fluorescence yields ; unhappily the theoretical calculations and the experimental determinations of R_K and R_L are not accurate enough. The fluorescence yield curve of the K shell is given in figure 13 according to Burhop's semi-empirical formula (1955)

$$\left(\frac{R_K}{1-R_K}\right)^{1/4} = B(Z-\sigma) + CZ^3. \qquad \ldots\ldots(13)$$

This relation takes screening into account by the σ parameter, and also relativistic effects which are important for high Z ; this is in agreement with the recent experimental data (Laberrigue-Frolow and Radvanyi 1956, Drever and Moljk 1957, Fairbrother, Parkyn and O'Connor 1957).

The L level fluorescence yields are known with still less accuracy. The L_{II} and L_{III} yields are important in K capture, the L_I fluorescence yield especially in

L capture (2s). In this last case, a supplementary Auger effect can occur, being due to the very low energy electrons emitted by Coster–Kronig transitions. The most recent analysis of L fluorescence yields has been made by Burhop (1955). In figure 14 curves have been drawn for $R(L_I)$, $R(L_{II})$, $R(L_{III})$ according to

$$\left(\frac{R_L}{1-R_L}\right)^{1/4} = B(Z-\sigma) \qquad \ldots\ldots(14)$$

neglecting the relativistic correction term of the order Z^3.

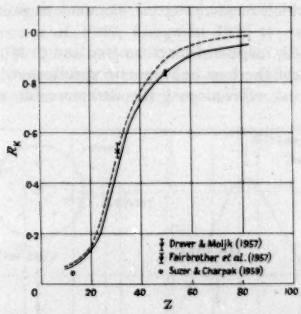


Figure 13. Fluorescence yield of the K shell according to the relativistic formula (13) with $B = 0{\cdot}0346$, $\sigma = 1{\cdot}27$, $C = -1{\cdot}35 \times 10^{-6}$. The non-relativistic curve (broken line) corresponds to $C = 0$.

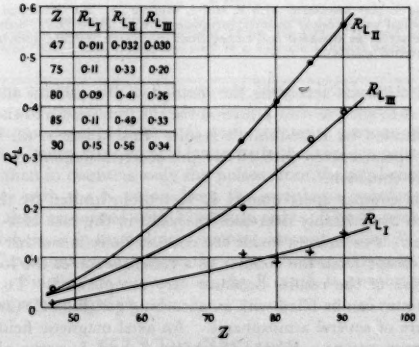


Figure 14. Fluorescence yields of the L shell from experimental data on line widths. (*After* Burhop 1955.) The $R(L_{III})$ curve can be represented by the non-relativistic relation

$$\{R(L_{III})/[1-R(L_{III})]\}^{1/4} \simeq B(Z-\sigma)$$

with $B = 0{\cdot}01080$ and the screening constant $\sigma \simeq 8$.

1.3.3. *Measurements of x-rays and Auger electrons.*

Energy and intensity measurements of x-rays and Auger electrons are very difficult; they have been carried out by several authors by means of Wilson chambers (Radvanyi 1954, 1955), Geiger–Müller counters (Alvarez 1937, Bouchez 1952), proportional counters (Pontecorvo *et al.* 1949, Avignon 1953, Langevin 1955 a), scintillation spectrometers with or without the radioactive element (Mateosian 1953), magnetic spectrometers (Haynes and Achor 1955, Bellicard and Moussa 1957) or even x-ray crystal spectrometers (Abelson 1939, Cook 1951).

We will not discuss here these different methods, but we want to draw attention to two methods, for which recent progress has made it possible greatly to improve their accuracy.

Of all these methods, magnetic spectrometry and proportional counter spectrometry are without doubt the best. Magnetic spectrometry is limited on the low-energy side by the weak efficiency of the detectors for soft electrons. But the

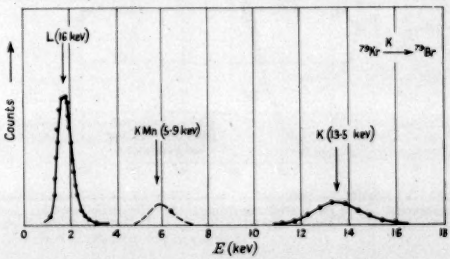


Figure 15. K and L rays associated with K and L capture by ^{79}Kr (Radvanyi 1955) and obtained with a classical proportional counter (propane 66·4 cm Hg + ^{79}Kr; diameter 8·4 cm). The 5·9 kev line serves as standard and comes from the K-ray of ^{55}Mn associated with K capture of ^{55}Fe (external source).

acceleration of electrons according the method of Rosenblum and Guillot (1937) should enable us to analyse with precision the electrons down to energies lower than 1 kev, as is indicated by Rosenblum's results (Rosenblum *et al.* 1956) by using a 180° magnetic spectrometer with 10 kev pre-acceleration and detection by means of photographic emulsion.

Proportional counter spectrometry is very often limited by the weak counter efficiency which considerably decreases the yield in the case of x-ray sources outside the counter. For sources inside the counter there is another difficulty. The radiation may escape from the counter as a consequence of the low efficiency and the interpretation of the results becomes very complicated. To overcome these defects, the counter can be filled with an absorbing gas (high Z) (krypton or xenon) under a pressure of several atmospheres. An axial magnetic field also enables us to reduce electron escape. Figure 15 shows K and L peaks obtained with an ordinary proportional counter and due to x-rays and soft electrons associated with K and L captures of ^{79}Kr. The K capture line is due partly to the K x-rays absorbed by the gas in the counter, relative intensity $KR_K(1-f)$ (where K is the

probability of K capture, f the probability of escape for a K photon), and partly to Auger electrons and very soft accompanying x-rays entirely absorbed in the counter intensity $K(1-R_K)$. The L line is due not only to radiations associated with L capture, but also to the soft radiations associated with the K_α photons which are not absorbed in the gas: intensity KR_Kfk (where k is the proportion of the K_α photons in the K series). The ratio of the areas of these peaks L and K is

$$\frac{N_L}{N_K}=\frac{L/K+R_K kf}{1-R_K f}. \qquad \ldots\ldots(15)$$

If the experimental conditions are such that no photon can escape from the counter, one has $f=0$ (Pontecorvo *et al.* 1949, Langevin 1955 a); and N_L/N_K is a direct measurement of the L/K branching ratio. However, when the energy of the K photon increases, this condition is more and more difficult to fulfil, the filling pressure or the counter diameter becoming prohibitive.

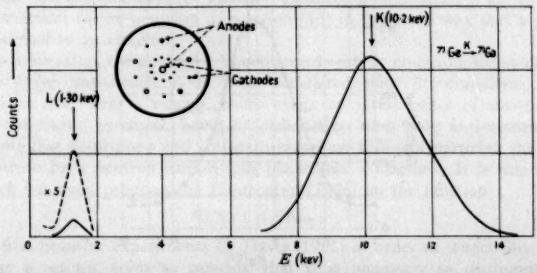


Figure 16. L and K lines coming from the central proportional counter, in anticoincidence with the annular counter (argon 6 atm + methane 15 cm Hg + ^{71}Ge ^{3}H) and associated with K and L capture in ^{71}Ge. The L line, due only to L capture, is much weaker than in figure 15.

Another recent method, due to Drever and Moljk (1957), enables us in certain cases to avoid corrections due to radiation escape and fluorescence yield. It uses a double proportional counter (figure 16) with an intense gaseous source, the axial measuring counter being in anticoincidence with the annular counter. Drever's method consists then in counting only the pulses from the axial counter which are in anticoincidence with the K pulses from the peripheric counter; the intensity of the L peak is then associated with L capture only. The K peak is due in part to the events which occur in the ordinary counter, intensity $K(1-R_K)+KR_K(1-f)$, and in part to K photons coming from the outer counter, intensity approximately equal to KR_Kf. Finally the intensity of the K peak is proportional that of K capture, to a first approximation.

§ 2. The Neutrino

We sum up here briefly the characteristics of the neutrino, in relation to electron capture, taking into account the recent experiments made in 1957 on the non-conservation of parity, which show new and fundamental properties of this particle.

Up to 1956 all the data accumulated about the neutrino, both in β radioactivity and in meson physics, are analysed in the excellent survey paper by Ridley (1956).

2.1. '*Old*' *Properties of the Neutrino*

The neutrino hypothesis was put forward by Pauli (1933 a) at the seventh Solvay conference to enable the maintenance of the conservation laws of energy and spin in β radioactivity. Soon afterwards this hypothesis had its first success in the Fermi theory of β emission (1934), in which the neutrino plays an essential role.

Since the capture of an atomic electron by the nucleus is analogous to β^+ emission, it was natural also to apply the neutrino hypothesis to capture. But in capture the available energy is shared between the recoil nucleus and the neutrino, which carries away a definite energy. Therefore, the measurement of

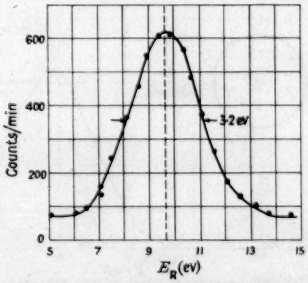


Figure 17. Energy spectrum of the recoil ions ^{37}Cl in the K capture of ^{37}A ; experimental result : (9·63 ± 0·06) ev ; theoretical result : (9·65 ± 0·05) ev corresponding to the emission of only one neutrino.

the recoil energy of the final nucleus is a good test of the validity of the neutrino hypothesis. The investigation of nuclear recoil energies in K capture has received much attention since its discovery by Alvarez *et al.* (1941). A remarkable experiment is that of Snell and Pleasonton (1955) on ^{37}A, which shows that the energy spectrum of the recoil nucleus is a line (figure 17) with an energy (9·63 ± 0·06) ev corresponding to the emission of only one particle, the mass of which is nearly zero.

Spin conservation in capture leads us, as in β emission, to attribute spin $\frac{1}{2}$ to the neutrino, this particle obeying Fermi–Dirac statistics.

The neutrino is evidently electrically neutral, to permit charge conservation in the transition. Its magnetic moment has not yet been detected ; recently Cowan and his associates (1954, 1956) have tried to measure the magnetic interaction between electrons in matter (liquid scintillator) and the neutrinos from a reactor : assuming the supplementary background noise when the reactor is present to be due entirely to the neutrinos, Cowan finds a cross section smaller than 7×10^{-16} barn, corresponding to a lower limit for the magnetic moment of the neutrino equal to 10^{-7} Bohr magneton.

Up to these very last years, the neutrino appeared as a still hypothetical particle, and the experiments on recoil nuclei in K capture, the strongest indication of its existence, provided still only an indirect proof.

But the fundamental experiments on the nuclear reactions initiated by the neutrino are a direct proof of its existence. The first nuclear reaction initiated by the neutrino has been carried out by Reines and Cowan (1956) who demonstrated the capture of a neutrino by a proton

$$p + \tilde{\nu} \rightarrow n + e^+ \qquad \ldots\ldots(16)$$

and measured a cross section of $(6 \pm 0{\cdot}6)\,10^{-20}$ barn. The protons (liquid scintillator) were exposed to an intense flux of neutrinos (about 10^{11} cm^{-2} sec^{-1}) from a reactor. This reaction (16) is the reverse of β^- emission

$$n \rightarrow p + e^- + \tilde{\nu} \qquad \ldots\ldots(17)$$

where the emitted neutrino is analogous to the one captured in the Cowan experiment (16). On the contrary, the neutrino captured by a neutron is not the same ; it is symmetrical to the neutrino in reaction (16) in the same way that a positron is symmetrical to an electron.

If, by convention, the neutrino ν is connected with β^+ emission, the symmetrical neutrino $\tilde{\nu}$ (or anti-neutrino) is then associated with β^- emission. It is a consequence that, near a reactor, there exists an intense flux of anti-neutrinos because the fission products, being β^- radioactive, emit only anti-neutrinos. If we suppose that neutrinos ν and anti-neutrinos $\tilde{\nu}$ are different particles, the capture of a neutrino by a neutron cannot take place near a reactor. It is this negative test which has been proposed by Pontecorvo (1946) on the reaction

$$^{37}Cl + \nu \rightarrow {}^{37}A + e^-, \qquad \ldots\ldots(18)$$

and the first negative experiments by Davis (1956) in order to detect this reaction (18) near a reactor seem to indicate that it is necessary to distinguish anti-neutrinos from β^- decay and neutrinos from β^+ decay.

The double β disintegration is another test of the non-identity of the neutrinos ; in fact, this reaction is much more probable for identical neutrinos than for different neutrinos. The experimental results (Awscholom 1956) on the $^{48}Ca \rightarrow Ti$ transition (figure 4 (*b*)) and on $^{150}Nd \rightarrow Sm$ (Cowan *et al.* 1956) are not positive and give only a lower limit of 2×10^{18} years for the half-life of the double β decay. On the contrary, the hypothesis of the identity of a neutrino and an anti-neutrino (Majorana 1937) would give for the double β decay of ^{48}Ca or ^{150}Nd a half-life shorter than about 10^{17} years.

With this old concept of the neutrino, the experiments on double β decay and neutrino capture seem to be against the Majorana hypothesis. Two different particles should exist, the neutrino connected with β^+ emission and electron capture, and the anti-neutrino connected with β^- emission.

2.2. '*New*' *Properties of the Neutrino*

Such were the ideas on the neutrino in 1956 before the upset caused by the anomalous K meson decay and its interpretation by Lee and Yang (1956). We do not wish to develop this question here. We simply recall that Lee and Yang introduced the principle of parity non-conservation to avoid the existence of two

K mesons ($K\pi_3$ or θ and $K\pi_3$ or τ) having the same mass, the same half-life but different parities. Lee and Yang generalized this parity non-conservation not only to β decay but to all weak interactions ; and all the experiments suggested by Lee and Yang have fully confirmed that the weak interactions do not conserve parity.

As to β interaction and the neutrino, the essential experimental facts concern (*a*) the asymmetry of the electrons emitted by oriented nuclei (Wu, Ambler and Hayward 1957) or by polarized μ mesons (Garwin, Lederman and Weinrich 1957, Friedman and Telegdi 1957, Muklin, Ozerov and Pontecorvo 1958), (*b*) the polarization of μ mesons coming from the disintegration of π mesons (Culligan *et al.* 1957), (*c*) the polarization of β decay electrons (Frauenfelder *et al.* 1957), (*d*) circular polarization of γ following β emission (angular correlation) (Boehm and Wapstra 1957, Schopper 1957), (*e*) circular polarization of γ following a K capture, in relation to the direction of the recoil nucleus (Goldhaber, Grodzins and Sunyar 1958).

All these results can be interpreted if one supposes that in every reaction with β decay the neutrino is emitted with a definite longitudinal polarization or ' helicity '.

The anti-neutrino appearing in β^- decay should have right-handed polarization (positive helicity). On the contrary, the neutrino appearing in β^+ decay or K capture should have a left-handed polarization (negative helicity). The neutrino can be represented as a particle with spin antiparallel to momentum (left-handed neutrino) and is connected with β^+ emission and K capture. The anti-neutrino has spin parallel to momentum (right-handed neutrino) and is connected with β^- decay. So the neutrino defines a left-handed system and the anti-neutrino a right-handed system. These particles are therefore responsible for the breakdown of parity conservation in β interaction.

Let us note that among the above experiments, that of Goldhaber *et al.* (1958) on K capture of ^{152}Eu is one of the most typical and gives a very clear indication of the helicity of the neutrino. Electron capture is in fact very convenient for studying the degree of polarization of the neutrino. The available energy from the transition is shared between the neutrino and the recoil nucleus, with the result that a polarization of the neutrino causes a polarization of the recoil nucleus ; and if this latter is excited with a very short half-life, it can, when returning to its ground state, emit a γ-ray which is generally circularly polarized. If one measures the polarization of this γ-ray, one can get the polarization of the neutrino. The first experimental result of this beautiful experiment is in agreement with a negative helicity for the neutrino emitted in electron capture. We remark further that the non-conservation of parity in the weak interactions other than β decay (π meson decay, K decay, Λ and Σ hyperon decays) indicates in every weak interaction either the presence of β interaction or rather the presence of an asymmetric type of coupling (Feynman and Gell-Mann 1958) which does not conserve parity.

2.3. *Asymmetry of the Neutrino*

2.3.1. *Dirac theory with four components.*

We suppose the Dirac theory to be applicable to any particle with spin $\frac{1}{2}$; and the neutrino can be represented by a Dirac four-component spinor

$$\phi = \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix} \qquad \ldots\ldots(19)$$

satisfying the Dirac equation without mass term

$$\gamma^\mu p_\mu \phi = 0 \qquad \ldots\ldots(20)$$

u and **v** being Pauli two-component spinors. We choose the matrix representation used by Dirac and Weisskopf

$$\alpha_k = \begin{pmatrix} \sigma_k \\ \sigma_k \end{pmatrix} = i\gamma^4\gamma^k \quad \text{where } \sigma_k \text{ are the Pauli matrices}$$

$$\beta = \gamma^4 = \begin{pmatrix} I \\ & -I \end{pmatrix}, \quad \text{where} \quad I = \begin{pmatrix} 1 \\ & 1 \end{pmatrix} \quad \text{and} \quad \gamma^5 = \gamma^1\gamma^2\gamma^3\gamma^4 = \begin{pmatrix} & -I \\ -I & \end{pmatrix}$$

with $k = 1$, 2 or 3 and $\mu = 1$, 2, 3 or 4.

It is well known that if the neutrino is described by a Dirac spinor, there exist four independent spinors ϕ_μ corresponding to four types of neutrino, the general solution being a linear combination of these spinors $\phi = \Sigma_\mu \alpha_\mu \phi_\mu$. And if the OZ axis is chosen along the direction of motion, each spinor takes the form

$$\phi_\mu = \frac{1}{\sqrt{2}} \eta_\mu e^{iqz} \qquad \ldots\ldots(21)$$

q being the positive momentum along OZ, $q = +\sqrt{(p^2)} = E_p$, η_μ the Dirac basic spinors defined in table 2. The positive energy solutions η_1 and η_2 being associated with the values $(E_p, \mathbf{p}, \boldsymbol{\sigma})$, and the negative energy solutions η_3 and η_4 with the values $(-E_p, -\mathbf{p}, -\boldsymbol{\sigma})$.

2.3.2. *Pauli two-component theory.*

As early as 1933 Pauli had already considered the non-conservation of parity for particles of zero rest mass obeying the Dirac equation. But he had rejected this result as inconsistent with physical reality. We know that this asymmetry must now be accepted.

The Dirac equation (20) without mass term splits into two equations

$$\boldsymbol{\sigma}\mathbf{p}\mathbf{v} = E\mathbf{u}, \quad \boldsymbol{\sigma}\mathbf{p}\mathbf{u} = E\mathbf{v}. \qquad \ldots\ldots(22)$$

For the neutrino moving with velocity c, the expectation value of the β operator is zero:

$$\langle\phi|\beta|\phi\rangle = \pm(1 - v^2/c^2)^{1/2} = 0$$

whence $|\mathbf{u}|^2 = |\mathbf{v}|^2$ and the two basic solutions are $\mathbf{v} = \mathbf{u}$ and $\mathbf{v} = -\mathbf{u}$.

(*a*) if $\mathbf{v} = \mathbf{u}$ the two equations (22) reduce to the unique equation

$$(\boldsymbol{\sigma}\mathbf{p} - E)\mathbf{u} = 0 \qquad \ldots\ldots(23)$$

considered by Pauli (1933 b), and again in 1957 by Lee and Yang, Salam (1957) and Landau (1957).

(*b*) the alternative solution $\mathbf{v} = -\mathbf{u}$ leads to the complementary equation

$$(\boldsymbol{\sigma}\mathbf{p} + E)\mathbf{u} = 0. \qquad \ldots\ldots(24)$$

Let us note that if one of the transformations $C = \gamma^2 K$ (charge conjugation) or $P = \gamma^4$ (parity) is applied to the equations (22), these two-component equations (23) and (24) transform into each other. On the contrary, they are invariant in the Lorentz transformation L, in the spatial rotation R and in the time inversion

$T = i\sigma_2 K$. The Dirac equation (20) is invariant in all these transformations (L, R, C, T and P).

2.3.3. *Parity, chirality and helicity.*

If the parity transformation $P = \gamma^4$ is applied to neutrinos $\mathbf{v} = \mathbf{u}$ described by (23) $\sigma p\mathbf{u} = E\mathbf{u}$, we obtain the two other neutrinos $\mathbf{v} = -\mathbf{u}$ described by (24) $\sigma p\mathbf{u} = -E\mathbf{u}$; and this separation can also be carried out from the Dirac spinor ϕ by the means of the projection operators $\frac{1}{2}(1 \pm \gamma^5)$: the Dirac function

$$\phi_- \equiv \tfrac{1}{2}(1-\gamma^5)\phi = \frac{1}{\sqrt{2}}(a_1\eta_1 + a_3\eta_3)\,e^{iqz} \qquad \ldots\ldots(25)$$

representing neutrinos $\mathbf{v} = \mathbf{u}$, and the other function

$$\phi_+ \equiv \tfrac{1}{2}(1+\gamma^5)\phi = \frac{1}{\sqrt{2}}(a_2\eta_2 + a_4\eta_4)\,e^{iqz} \qquad \ldots\ldots(26)$$

giving the two other neutrinos $\mathbf{v} = -\mathbf{u}$.

These functions ϕ_+ and ϕ_- are the eigenfunctions of the γ^5 operator and have opposite chirality

$$\begin{aligned} \gamma^5\phi_+ &= +\phi_+ \quad \text{with} \quad \langle\phi_+|\gamma^5|\phi_+\rangle = +1 \\ \gamma^5\phi_- &= -\phi_- \quad \text{with} \quad \langle\phi_-|\gamma^5|\phi_-\rangle = -1. \end{aligned} \qquad \ldots\ldots(27)$$

They transform into each other by parity transformation, $\gamma^4\phi_+ = \phi_-$ and vice versa, as would right-handed and left-handed neutrino wave functions. For this reason, γ^5 is the chirality operator (from the Greek word ' cheir ' = hand); these functions ϕ_+ and ϕ_- also transform into each other by C charge conjugation.

Lee and Yang (1957) also introduced the longitudinal polarization or helicity $H = \langle\boldsymbol{\sigma}\mathbf{p}\rangle/q$; the positive energy solution $\eta_1 e^{iqz}$ has a positive helicity ($H = +1$) and $\eta_2 e^{iqz}$ a negative helicity ($H = -1$). We can also see that for negative energy solutions helicity H and chirality χ are the same, but opposite for positive energy solutions.

Table 2. Different Kinds of Neutrinos (the term e^{iqz} is neglected)

ϕ	η_1	η_2	η_3	η_4
$\mathbf{u}$	1	0	1	0
	0	1	0	1
$\mathbf{v}$	1	0	1	0
	0	−1	0	−1
E	$+E_p$		$-E_p$	
$H = \dfrac{\langle\boldsymbol{\sigma}\mathbf{p}\rangle}{q}$	1	−1	−1	+1
$\chi = \langle\gamma^5\rangle$	−1	+1	−1	+1
ν	ν_R	ν_L	$\bar{\nu}_R$	$\bar{\nu}_L$

2.3.4. *Four or two neutrinos.*

In the old experiments there was no reason to choose one kind of neutrino rather than another; it was assumed that the four types of neutrino, like the electrons, existed in nature: two neutrinos and two anti-neutrinos.

However, Majorana, in 1937, had supposed that neutrinos and anti-neutrinos were identical, since these particles have no charge and no magnetic moment; this meant that in nature there are only two neutrinos. But before the new experiments in 1956, this hypothesis was rejected because the results on double β decay and neutrino capture by ^{37}Cl seemed to be against this point of view.

Now, the evidence for non-conservation of parity in weak interactions leads us to assume again the existence of two neutrinos.

2.3.5. *Lepton conservation.*

The experiments on parity non-conservation show that in every type of β decay a neutrino with definite helicity is emitted. This spatial asymmetry, due to the helicity of the neutrino, is the reason for non-conservation of parity. The sign of this helicity was determined after the recent experiments on the sign of the polarization of the μ meson in π decay (Culligan *et al.* 1957), and on the sign of the circular polarization of γ rays following K capture (Goldhaber *et al.* 1958): in β^- and π^- decay, e^- and μ^- are accompanied by right-handed neutrinos ($H = +1$) and in β^+ and π^+ decay, by left-handed neutrinos ($H = -1$).

If we suppose that e^- (or μ^-) is a particle and e^+ (or μ^+) an anti-particle, one can consider that the left-handed neutrino ν_L accompanying e^+ (or μ^+) is a particle, and the right-handed neutrino accompanying e^- (or μ^-) is the anti-particle $\tilde{\nu}_L$.

We can then introduce a lepton charge l for light particles: $l = +1$ for particles, $l = -1$ for anti-particles ($l = 0$ for heavy particles π, n, p...). It seems that $\Sigma l_i = 0$ in transitions with weak interactions; this hypothesis is the 'lepton conservation'; the number l satisfies a law of conservation like the magnetic quantum number.

2.3.6. *The two-component theory.*

These experimental results do not, however, prove the existence of two neutrinos instead of four. In fact, there are two neutrinos having the same helicity and it could be assumed that in every type of β decay, corresponding to neutrinos of definite helicity, sometimes the neutrino is emitted and sometimes the anti-neutrino with the same helicity. For instance, in K capture we should have

$$\begin{aligned} &(p) + e^- \rightarrow (n) + \nu_L \quad (H = -1) \\ &(p) + e^- \rightarrow (n) + \tilde{\nu}_R \quad (H = -1). \end{aligned} \qquad \ldots\ldots(28)$$

But this hypothesis is not attractive and for the sake of simplicity one can retain the opposite point of view: in each β decay only one kind of neutrino is emitted

$$\begin{aligned} &(p) + e^- \rightarrow (n) + \nu_L \quad e^- \text{ capture} \\ &(p) \rightarrow (n) + e^+ + \nu_L \quad \beta^+ \text{ emission} \\ &(n) \rightarrow (p) + e^- + \tilde{\nu}_L \quad \beta^- \text{ emission.} \end{aligned} \qquad \ldots\ldots(29)$$

This point of view is equivalent to assuming that there exist in nature only two neutrinos, one neutrino ν_L and one anti-neutrino $\tilde{\nu}_L$; these two neutrinos obey the two-solution equation $(\boldsymbol{\sigma}\mathbf{p} + E)\mathbf{u} = 0$: this is the two-component theory.

2.3.7. *The new Majorana hypothesis.*

Another phenomenological interpretation is possible. We observe only the left- or right-handed character, but not the particle or anti-particle character. The old Majorana hypothesis can then be applied to the left or right neutrino: the right neutrino is identical with the anti-left neutrino, $\nu_R = \tilde{\nu}_L$, the left neutrino is identical with the anti-right neutrino, $\nu_L = \tilde{\nu}_R$.

We have again two kinds of neutrinos, the left-handed and the right-handed one. The left-handed neutrino is associated with the right-handed e^+ (or μ^+), and the right-handed neutrino with the left-handed e^- (or μ^-).

The double β decay corresponds then to

$$(n\text{–}n) \rightarrow (p\text{–}p) + (e_L^- + \nu_R) + (e_L^- + \nu_R) \qquad \ldots\ldots(30)$$

and four particles are emitted, which means a very long life.

This point of view corresponds to the γ^5 invariance or chirality conservation taking the place of lepton conservation.

§ 3. Fermi Interaction and Orbital Capture

The theory of electron capture has been worked out in the same way as the theory of β emission due to Fermi (1934). In his original paper on " Search for a theory of β radiation ", Fermi chooses a nucleon–lepton interaction such that " a β^- transition should correspond to the creation of an electron and a neutrino ". Taking the Dirac theory into account, where the creation of a particle is equivalent to the annihilation of its anti-particle, Fermi considers that " a β^+ transition corresponds to the annihilation of an electron and a neutrino ".

With this formulation, the Fermi theory of β emission included in an implicit manner the capture of an electron with emission of an anti-neutrino. This extension of the β-decay theory to orbital capture was made independently by Yukawa and Sakata (1935), Bethe and Bacher (1936) and Møller (1937 a, b). These authors had calculated the transition probabilities of K capture before its discovery.

The success of this theory indicates that electron capture must be considered as a particular case of β decay, in the same way as β emission, μ decay or μ^- capture. All these processes should be produced by the same unique interaction between heavy and light particles: the nucleon–lepton interaction or Fermi interaction.

The Fermi interaction is not yet very accurately known, but recent experiments have made possible much progress in its understanding. It can be said that the main purpose of the experiments in the field of β decay is the determination of the Fermi interaction, that is to say the determination of the coupling constants as well as the nature of the neutrino.

We have seen (29) that the very recent experiments have made it possible to interpret electron capture as

$$(p) + e^- \rightarrow (n) + \nu_L \qquad \ldots\ldots(31)$$

with emission of a left-handed neutrino which in Dirac theory is represented by the spinor $\frac{1}{2}(1+\gamma^5)\phi$.

We can draw some conclusions concerning the Fermi interaction. The Fermi interaction between particles in electron capture is described by a Hamiltonian H

analogous to that of β emission, and the interaction energy

$$\langle H \rangle = \int J(\mathbf{r})\, d^3(\mathbf{r}) \qquad \ldots\ldots(32)$$

where $J(\mathbf{r})$ is the product of particle functions taken at the same point $\mathbf{r}$, has in the ' old ' theory the general form

$$J = g \sum_{\lambda} C_\lambda (\psi_n^+ O_\lambda \psi_p)(\phi^+ O_\lambda \psi) \qquad \ldots\ldots(33)$$

$\psi_p = |E, IM\pi\rangle$ being the initial nuclear state with eigenvalues E (energy), I and M for angular momentum, π for parity ; $\psi_n = |E', I'M'\pi'\rangle$ is the final nuclear state ; $\psi = |\epsilon, j_1 m_1\rangle$ the bound electron in the initial atom Z and having a total energy less than $m_0 c^2$ $(\epsilon < 1)$; $\phi = |q, j_2 m_2\rangle$ being a neutrino state with kinetic energy $q > 0$, free particle in an unbound state (continuous energy spectrum), the operator Q $(Q\psi_p = \psi_n)$ making the transition from a proton state to a neutron state being omitted.

This density J is a Lorentz invariant linear combination of the five Lorentz invariants which can be built from the Dirac spinors $\psi_p \psi_n \psi \phi$ by the multiplication of two scalars S, two vectors V, two tensors T, two axial-vectors A or two pseudo-scalars P.

Using the γ matrices, the operators O_λ can be written

Scalar	$O_S = \gamma^4$	
Vector	$O_V = \gamma^4 \gamma^\mu$	
Tensor	$O_T = -i\gamma^4 \gamma^\mu \gamma^\nu$	$\ldots\ldots(34)$
Axial-vector	$O_A = i\gamma^4 \gamma^5 \gamma^\mu$	
Pseudo-scalar	$O_P = \gamma^4 \gamma^5$	

where μ and ν run from 1 to 4. Using Dirac notation, the operators O_λ are identical with the expressions of Konopinski (Konopinski and Uhlenbeck 1941, Konopinski 1943)

$$\begin{aligned} O_S &= \beta & O_A &= \sigma - i\gamma^5 \\ O_V &= 1 - i\alpha & O_P &= \beta\gamma^5 \\ O_T &= \beta\sigma + \beta\alpha \end{aligned} \qquad \ldots\ldots(35)$$

In this old theory, the Hamiltonian is γ^4 invariant (parity conservation), the neutrino is emitted in the right-handed state ν_R as well as in the left-handed state ν_L and parity is evidently conserved.

The new experiments show on the contrary that, in capture, the neutrino is only emitted (to a good approximation) in the left-handed state ν_L, the consequence being non-conservation of parity. Let us note, however, that the present experimental results (Wu 1959) allow only the conclusion that the left-handed neutrino is emitted with a frequency of 80%. To be quite rigorous, one should write for electron capture (or β^+ emission)

$$C_\lambda \phi \to C_\lambda^L \phi_L + C_\lambda^R \phi_R \quad \text{with} \quad |C_\lambda^R| \ll |C_\lambda^L| \qquad \ldots\ldots(36)$$

the interaction constants C_λ^{L} and C_λ^{R} taking into account the proportions of left and right neutrinos which are emitted. But for the sake of simplicity we shall assume that only left-handed neutrinos are emitted.

In order to determine the interaction constants more completely we have to consider other experimental results. We can easily discuss the polarization of the electrons emitted in β decay. This result on polarization leads to the conclusion that only (V, A) interactions are effective, and not (S, T) interactions (we can exclude the P interaction for the moment).

We can reach this important conclusion by using the fact that the positive electrons are emitted with a $+v/c$ maximum longitudinal polarization. If these electrons were very fast, the degree of polarization would be equal to unity $(+1)$ and we can say that the electrons have in β^+ decay a right-handed character. The β^+ decay

$$(\mathrm{p}) \rightarrow (\mathrm{n}) + \mathrm{e}_{\mathrm{R}}{}^{+} + \nu_{\mathrm{L}} \qquad \ldots\ldots(37)$$

corresponds to the emission of a right-handed e^+ and a left-handed ν (or also theoretically to the capture of a right-handed anti-neutrino).

The lepton term of the interaction (33) can be written (if we omit the numerical factor)

$$\psi^+ O_\lambda \phi_{\mathrm{R}} = \psi^+ O_\lambda (1-\gamma^5)\phi \qquad \ldots\ldots(38a)$$

which transforms mathematically in two different ways according to the choice of O_λ operators :

odd operators O_1, with an odd number of γ^μ matrices (S or T)

$$\psi^+(1+\gamma^5) O_1 (1-\gamma^5)\phi = \psi_{\mathrm{L}}{}^{+} O_1 \phi_{\mathrm{R}} \qquad \ldots\ldots(38b)$$

even operators O_2, with an even number of γ^μ matrices (V or A)

$$\psi^+(1-\gamma^5) O_2 (1-\gamma^5)\phi = \psi_{\mathrm{R}}{}^{+} O_2 \phi_{\mathrm{R}}. \qquad \ldots\ldots(38c)$$

It is evident that only the (V, A) interaction gives for the emitted positron a right-handed character, which is consistent with polarization experiments.

To be consistent with the helicity of the neutrino and the maximum polarization of the electrons in β emission, the ' old ' interaction (33) for electron capture (or β^+ emission) must be written with simplified notations

$$J = g[2C_{\mathrm{V}}(\mathrm{n}^+, \mathrm{V}, \mathrm{p})(\nu_{\mathrm{L}}{}^{+}, \mathrm{V}, \mathrm{e}_{\mathrm{L}}) + 2C_{\mathrm{A}}(\mathrm{n}^+, \mathrm{A}, \mathrm{p})(\nu_{\mathrm{L}}{}^{+}, \mathrm{A}, \mathrm{e}_{\mathrm{L}})] \qquad \ldots\ldots(39)$$

C_{V} and C_{A} being normalized so that $|C_{\mathrm{V}}|^2 + |C_{\mathrm{A}}|^2 = 1$, and the factor 2 before C_{V} and C_{A} giving for the interaction constant g the same value as the old definition (33).

We have still to get a precise value of the coefficients C_{V} and C_{A} from other experiments. On one hand, according to Liubimov's (1959) experiments on the determination of the neutron half-life the ratio of V and A coupling constants is $|C_{\mathrm{A}}|^2/|C_{\mathrm{V}}|^2 = 1{\cdot}55 \pm 0{\cdot}08$. On the other hand the experiments on the β decay of polarized neutrons (Burgy *et al.* 1959) seem to indicate real coefficients C_{A} and C_{V} (T invariance) and a predominantly (V, A) coupling.

In the following, if not specified, the C_{V} and C_{A} coefficients will be taken as arbitrary constants.

§ 4. Transition Probabilities for Capture

The measurement of a scalar evidently does not give any evidence for parity non-conservation: transition probability, β spectrum shape, correlation β recoil, branching ratio K/L or K/β^+. Consequently the asymmetry of the neutrino will not modify the transition probabilities and the branching ratios. But since the β interaction is now (V, A) and not (S, T), the forbidden transition probabilities will be changed.

In this paper, we shall give the theory of electron capture, with the (V, A) interaction and use the method of irreducible tensors.

Let us recall that the theory of capture was first made by Yukawa and Sakata (1935), Bethe and Bacher (1936) and Møller (1937) in the case of allowed transitions. Calculations for forbidden transitions have been carried out by Marshak (1942), Bouchez (1950), Nataf and Bouchez (1952) and lately by Brysk and Rose (1955) for (S, T) coupling. The new results for (V, A) coupling by Brysk and Rose (1958) will be compared with these.

4.1. *General Expression of the Transition Probabilities*

The general expression for the transition probability per unit of time, between the initial state ψ_p and the final state ψ_n of the nucleus, is given by the perturbation theory. Limited to the first order, the probability is

$$P = 2\pi \Sigma |\langle \psi_n | \mathscr{H} | \psi_p \rangle|^2 \qquad \ldots\ldots(40)$$

the units being m_0, c and $\hbar$, the functions normalized to one particle per unit energy interval, the sum to be carried out over the M' final states of the same energy and corresponding to the different states $|j_1 m_1\rangle$ of the captured electron and $|j_2 m_2\rangle$ of the emitted neutrino, consistent with the transition $|IM\rangle \to |I'M'\rangle$ and also the average over all the M initial states.

The transition operator $\mathscr{H}$, or the energy density $J = \psi_n^+ \mathscr{H} \psi_p$ can be written after (39), using tensor notation

$$J = 2g \sum_{\rho=1}^{4} (-1)^{\rho+1} C_\rho \sum_{\mu=-1}^{1} (\psi_n^+ \omega_\rho^{-\mu} \psi_p)(-1)^\mu (\phi^+ \omega_\rho^\mu \psi) . \delta(\Omega_\lambda - \Omega_\nu) \qquad \ldots\ldots(41)$$

where:

$$\left.\begin{array}{llllll} \text{for} & \rho = 1 & C_1 = C_V & \text{and} & \omega_1 = 1 \\ & \rho = 2 & C_2 = C_V & \text{and} & \omega_2 = \alpha \\ & \rho = 3 & C_3 = C_A & \text{and} & \omega_3 = \sigma \\ & \rho = 4 & C_4 = C_A & \text{and} & \omega_4 = \gamma^5 \end{array}\right\} \qquad \ldots\ldots(42)$$

and the scalar product $A.B$ is defined by

$$A.B = \sum_{\mu=-1}^{1} A^\mu (-1)^\mu B^{-\mu} \qquad \ldots\ldots(43)$$

for two vector operators or for two scalar operators ($\mu = 0$ only). The Dirac function $\delta(\Omega_\lambda - \Omega_\nu)$ describes the contact β interaction, on the sphere of radius R (nuclear radius), between the nucleon at the point $\Omega_\nu(\theta_\nu, \phi_\nu)$ and the lepton at the same point $\Omega_\lambda(\theta_\lambda, \phi_\lambda)$.

The matrix element $\langle \mathscr{H} \rangle$ for the transition becomes

$$2g \sum_{\rho} (-1)^{\rho+1} C_{\rho} \sum_{\mu} (-1)^{\mu} \langle \psi_n | \omega_{\rho}^{-\mu}, L(\omega_{\rho}^{\mu}) \delta(\Omega_{\lambda} - \Omega_{\nu}) | \psi_p \rangle \quad \ldots\ldots(44)$$

where $L(\omega) \equiv \phi^{+} \omega \psi$ is the lepton element.

The δ function can be developed into a series of $i^{l} Y_l$ spherical harmonic products

$$\delta(\Omega_{\lambda} - \Omega_{\nu}) = \sum_{l} (-1)^{l} \sum_{m} (-1)^{m} Y_l^{-m}(\Omega_{\nu}) Y_l^{m}(\Omega_{\lambda}) \quad \ldots\ldots(45)$$

and each matrix element can be written

$$2g \sum_{\rho \mu l m} (-1)^{m+\mu+l+\rho+1} C_{\rho} \langle I'M' | \omega_{\rho}^{-\mu} Y_l^{-m} | IM \rangle \langle j_2 m_2 | \omega_{\rho}^{\mu} Y_l^{m} | j_1 m_1 \rangle \quad \ldots\ldots(46)$$

and split into a nuclear element and a lepton matrix element.

Let us introduce the spherical tensor of rank k

$$T_k^{q}(\omega_{\rho} Y_l) = \sum_{m\mu} (l \xi m \mu | kq) Y_l^{m} \omega_{\rho}^{\mu} \quad \ldots\ldots(47)$$

with $\xi = 0$ for scalar operators 1 and γ^5, and $\xi = 1$ for vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}$; $(l \xi m \mu | kq)$ is a Clebsch–Gordan coefficient. This definition of spherical tensors is analogous to that used by Blatt and Weisskopf (1952), but differs from that used by Rose and Osborn (1954). From relation (47) we calculate the inverse relation

$$Y_l^{m} \omega_{\rho}^{\mu} = \sum_{kq} (l \xi m \mu | kq) T_k^{q}(\omega_{\rho} Y_l). \quad \ldots\ldots(48)$$

From relation (48) and the properties of Clebsch–Gordan coefficients, the density J becomes

$$J = 2g \sum_{\rho l} \sum_{kq} (-1)^{q+\xi+k+\rho+1} C_{\rho} (\psi_n^{+} T_k^{-q} \psi_p)(\phi^{+} T_k^{q} \psi). \quad \ldots\ldots(49a)$$

Hence the energy of interaction $\langle \mathscr{H} \rangle$ becomes, more briefly,

$$\langle \mathscr{H} \rangle = 2g \sum_{\rho l k} C_{\rho} (-1)^{k+\xi+\rho+1} \langle T_k \rangle \langle L_k \rangle \quad \ldots\ldots(49b)$$

where

$$\langle T_k \rangle \langle L_k \rangle = \sum_{q} (-1)^{q} \langle T_k^{-q} \rangle \langle L_k^{q} \rangle$$

is the scalar product of the matrix elements of the irreducible nuclear and lepton tensors.

The calculation of the lepton term $L(\omega_{\rho})$ is very easy in the two-component theory. We obtain

$$L(1) = -L(\gamma^5) = \mathbf{u}^{+} \mathbf{v} \quad \text{and} \quad L(\sigma_k) = -L(\alpha_k) = \mathbf{u}^{+} \sigma_k \mathbf{v}$$

σ_k being the Pauli 2×2 matrices and the Pauli spinors $\mathbf{u}$ and $\mathbf{v}$ being related to the Dirac spinors of neutrino and electron

$$\phi_L \equiv \tfrac{1}{2}(1+\gamma^5)\phi = \frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{u} \\ -\mathbf{u} \end{pmatrix} \quad \text{and} \quad \psi_L \equiv \tfrac{1}{2}(1+\gamma^5)\psi = \frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{v} \\ -\mathbf{v} \end{pmatrix}.$$

4.2. *Selection Rules*

The selection rules enable us to select the matrix elements which do not vanish in the transition $|IM\rangle \rightarrow |I'M'\rangle$. These rules are analogous to those for β emission and correspond to the conservation laws for total angular momentum I and parity π of the nuclear levels.

These conservation laws can be obtained by writing the various matrix elements in the following way, from (49),

$$\langle I'M'\pi'|\,T_{kl}{}^{-q}(\omega_\rho\,\mathrm{Y}_l)\,|IM\pi\rangle\langle j_2 m_2|\,T_{kl}{}^{q}(\omega_\rho\,\mathrm{Y}_l)|j_1 m_1\rangle \quad \ldots\ldots(50)$$

giving immediately for magnetic quantum numbers the conservation law

$$M+m_1 = M'+m_2.$$

Conservation of total angular momentum. The well-known relation between **j** vectors in orbital capture

$$\mathbf{I}+\mathbf{j}_1 = \mathbf{I}'+\mathbf{j}_2$$

becomes in the case of the matrix elements (49)

$$\begin{aligned} |I-I'| &\leqslant k \leqslant I+I' \\ |j_1-j_2| &\leqslant k \leqslant j_1+j_2. \end{aligned} \qquad \ldots\ldots(51)$$

The possible values of the k parameter are given by this relation (51) and enable us to select for each operator ω_ρ the l values of Y_l which give non-vanishing matrix elements. Among the possible numbers k, the smallest ones only are useful and lead to notable values for the matrix elements; the others, associated with large values of j_1 and j_2 correspond to terms in $(\alpha ZR)^{j_1}$ and $(qR)^{j_2}$ which are negligible, the neutrino being emitted with a wavelength very large compared with the nuclear radius ($\lambda \gg R$ or $qR \ll 1$).

Conservation of the parity of nuclear levels. The non-conservation of parity for the neutrino does not break the conservation of parity for nuclear levels. Parity remains a good quantum number for nuclear states and it follows that the nuclear matrix element (50) satisfies

$$\pi\,\pi' = \pi(T_{kl}) = (-1)^l \pi(\omega_\rho).$$

This condition eliminates the l indices of the spherical harmonics which are not compatible with the parity of the operator ω_ρ and with the parity change $\Delta\pi = \pi\,\pi'$ of the transition.

On the contrary, this condition is not applicable to the lepton matrix element; parity being indeterminate for a neutrino, it follows that the eigenfunction j_2 of the neutrino can be built from Dirac functions of the two types

$$j_2 = l_2+\tfrac{1}{2} \quad \text{and} \quad j_2 = (l_2+1)-\tfrac{1}{2}.$$

The parity of a relativistic electron may be defined somewhat arbitrarily as the parity of the ψ_3 and ψ_4 components of the Dirac spinor. When $v \ll c$, the electron being non-relativistic, the large components reduce to the Pauli spinor and the parity is defined by Y_l (s, p, d... states).

Conservation of orbital angular momentum. The orbital angular momentum L of a nuclear state is not a constant of motion and cannot characterize a nuclear level. However, it is always possible to expand a state $|IM\rangle$ in a series of configurations $|IML\rangle$; if the number of terms in this series is not too large, it is then convenient to characterize the level by L configurations. An ideal case is that of an odd A nucleus of the type $(O+1)$ with only one extra nucleon in an l state; this nucleus is in the state $L = l$. The nucleons in the nucleus being non-relativistic, the parity of this level is $(-1)^L$. In a transition between two odd A nuclei of this type, the

change of orbital angular momentum $L = l' - l$ characterizes the transition, and the change of parity, in particular, is given by ΔL.

For nuclei of the type $(O+2)$ with two extra nucleons (l_1 and l_2) the levels evidently are not pure L states. But if we limit ourselves to the non-relativistic approximation, and if the two extra nucleons attract one another strongly, the $L = l_1 - l_2$ and $l_1 - l_2 + 2$ configurations will be predominant, and the parity of the nuclear level can still be defined as $(-1)^L$. A transition between two levels of this type is described by a linear combination of matrix elements having the form

$$\sum_{LL'} C_{LL'} \langle J'L' | T_k | JL \rangle. \qquad \ldots\ldots (52)$$

In this case, the minimum value of $\Delta L = |L' - L|$ generally characterizes the transition, the value $\Delta L_{\min} + 2$ leading to a much smaller probability.

Table 3. Selection rules giving the T_k tensors of rank k appearing in the transitions characterized by ΔI and $\Delta\pi$, and leading to the most important matrix elements. The special forbidden transitions have been indicated in each case : for instance $0 \leftrightarrow 0$ for the operator $T_1(\sigma Y_0)$ meaning that this tensor cannot produce this transition.

ΔI	$\Delta\pi$	Y_0				Y_1			
		1	σ	γ^5	α	1	σ	γ^5	α
0	$+1$	T_0	$T_1(0\leftrightarrow 0)$				T_0		
	-1			T_0	$T_1(0\leftrightarrow 0)$	$T_1(0\leftrightarrow 0)$	$T_1(0\leftrightarrow 0)$		
							$T_2\left(\begin{smallmatrix}0\leftrightarrow 0\\ \frac{1}{2}\leftrightarrow\frac{1}{2}\end{smallmatrix}\right)$		
1	$+1$		T_1						
	-1				T_1	T_1	T_1		
							$T_2(0\leftrightarrow 1)$		

ΔI	$\Delta\pi$	Y_{n-1}				Y_n			
		1	σ	γ^5	α	1	σ	γ^5	α
n	$(-1)^n$				T_n	T_n	T_n		
	$(-1)^{n+1}$		T_n				$T_{n+1}(0\leftrightarrow n)$		

A remarkable case is that of l-forbidden transitions when the minimum value $\Delta L = 0$ is present in a very small proportion, but sufficiently strong to make this transition $\Delta L = 0$ predominant. Then, the transition has not the properties of an l-forbidden transition but of an allowed transition (spectrum shape, K/β^+ ratio), and its fT product only corresponds to that characteristic of a forbidden transition (Bouchez 1950, Pursey 1951, Nataf and Bouchez 1952). The nuclides ^{22}Na and ^{65}Zn are examples of this in the case of electron capture and β^+ emission, while the nuclides ^{14}C and ^{32}P are probably examples of it in the case of β^- emission.

The above selection rules indicate for every transition $(\Delta I, \Delta\pi)$ the principal terms which give appreciable contributions to the transition probability. In table 3 are shown the matrix elements associated with the different transitions. When these terms are classified, the magnitude of the lepton matrix elements is given by the radial Bessel functions of the neutrino $f_l^B, g_l^B, f_{-l-1}^B, g_{-l-1}^B$ and those of

the electron $f_K, g_K, f_{L_{II}}, g_{L_{II}}$ which are known with great accuracy. On the contrary, the value of the nuclear matrix elements is not known, and only the order of magnitude can be settled. If we consider only the (V, A) interaction, we can compare the matrix elements $\langle 1\rangle, \langle\sigma\rangle, \langle\alpha\rangle$, and $\langle\gamma^5\rangle$; $\langle 1\rangle \simeq \langle\sigma\rangle$ and

$$\langle\alpha\rangle \simeq \langle\gamma^5\rangle \simeq (\bar{v}/c)\langle 1\rangle,$$

$\bar{v}$ being the mean velocity in the nucleus of the nucleon which undergoes the transition : $\bar{v}/c \simeq 0{\cdot}2$.

4.3. *Nuclear Matrix Elements and $f_K T_K$ Products*

4.3.1. *Form of the matrix elements.*

The probability of transition (41) can generally be written, observing that the matrix elements split into nuclear and lepton elements (45),

$$P = 8\pi g^2 \sum_{M'} \sum_{j_1} \sum_{m_1 m_2} \Big| \sum_{\rho} C_\rho \sum_{kl} (-1)^{k+\xi+\rho+1} \langle T_k\rangle \langle L_k\rangle \Big|^2. \quad \ldots\ldots(53)$$

The summation over the magnetic quantum numbers (M, M') and (m_1, m_2) of the nuclei and the lepton can be carried out in a straightforward manner by specifying the states $|j_1 m_1\rangle$ and $|j_2 m_2\rangle$ consistent with each transition

$$|IM\rangle \rightarrow |I'M'\rangle$$

(Bouchez 1950, Nataf and Bouchez 1952). But this summation can be performed more easily by means of the reduced matrix elements (Rose and Osborn 1954).

The nuclear and lepton matrix elements can be written in reduced form, after the Wigner–Eckart theorem

$$\begin{aligned} \langle IM'|\, T_k^{-q}\,|IM\rangle &= (IkM-q\,|\,I'M')\langle I'\|\, T_k \,\|I\rangle \\ \langle j_2 m_2|\, T_k^{q}\,|j_1 m_1\rangle &= (j_1 k m_1 q\,|\,j_2 m_2)\langle j_2\|\, T_k \,\|j_1\rangle. \end{aligned} \quad \ldots\ldots(54)$$

Taking into account the orthogonality relations between Clebsch–Gordan coefficients, the summation over $m_1 m_2$, M' and average over M give for the probability of capture of an electron in the state $|j_1\rangle$, from only one initial nuclear state $|IM\rangle$ to all final states $|I'M'\rangle$

$$P = 8\pi g^2 \frac{2I'+1}{2I+1} \sum_{\rho\rho'} C_\rho C_{\rho'} \sum_{kl} (-1)^{\rho+\rho'+\xi+\xi'} \sum_k \frac{1}{2k+1}$$

$$\times \sum_{j_2} (2j_2+1) \langle\|\, T_k(\omega_\rho \Upsilon_l)\,\|\rangle \langle\|\, T_k(\omega_{\rho'} \Upsilon_{l'})\,\|\rangle^* \langle\|\, L_k(\omega_\rho \Upsilon_l)\,\|\rangle \langle\|\, L_k(\omega_{\rho'} \Upsilon_{l'})\,\|\rangle^*. \quad \ldots\ldots(55)$$

There is no sum over j_1, the electron being captured in the definite state j_1. But the occupation $n(j_1)$ of a sub-shell j_1 can also be specified (Brysk and Rose 1955) by multiplying each term in (55) by $n(j_1)$ the relative occupancy of the sub-shell.

We will use the following notation for the nuclear terms

$$\frac{2I'+1}{2I+1} \langle\|\, T_k^{\rho l}\,\|\rangle \langle\|\, T_k^{\rho' l'}\,\|\rangle^* \equiv (\Upsilon_0^0)^2 M_k(\omega_\rho \Upsilon_l) M_k(\omega_{\rho'} \Upsilon_{l'})^* \quad \ldots\ldots(56)$$

the normalization factor $(\Upsilon_0^0)^2$ being introduced for a comparison with Møller's historic formula.

4.3.2. $f_K T_K$ *products.*

The above nuclear matrix elements cannot be calculated with accuracy. The direct calculation would require a complete understanding not only of the β interaction but also of the schemes of coupling (j–j, L–S) of the nucleons and our knowledge about nuclei is much too slight to permit such a calculation. But we can get the values of the matrix elements indirectly, and this leads to predictions about the coupling of nucleons and the structure of nuclei.

The determination of the squared nuclear matrix elements $|M|^2$ is achieved, as in β emission, by the means of the fT product which can be obtained from the capture transition probability

$$P_K = \frac{g^2}{4\pi^2}|M|^2 f_K \qquad \ldots\ldots(57)$$

if the other captures are neglected. The function f_K is given by the theory and contains only terms depending on the transition energy and the eigenfunctions of the electron and neutrino. All these terms are known very exactly, assuming only that the general theory of capture, i.e. the theory of β decay, is satisfactory. The partial probability P_K relative to the transition $|I\pi\rangle \rightarrow |I'\pi'\rangle$ is given by the determination of the partial period T_K, if the decay scheme is well known. The nuclear matrix element can be obtained from the product

$$f_K T_K = \frac{4\pi^2 L2}{g^2}\frac{1}{|M|^2}. \qquad \ldots\ldots(58)$$

The g value of the β interaction constant is very well known. For instance, it follows from the β^- disintegration of the neutron ($W_0 = 782$ kev, $T = 11{\cdot}3$ min) that

$$f_0 T = \frac{2\pi^2 L2}{g^2}\frac{1}{|M_0|^2} \quad \text{with} \quad f_0 \simeq \int_1^{\epsilon_0} \epsilon p q^2 F_0 \, d\epsilon \simeq 1{\cdot}72 \,\dagger$$

the nuclear matrix element $|M_0|^2$ of the neutron decay being assumed equal to unity. We obtain then for the neutron $f_0 T = T_0/|M_0|^2 \simeq T_0 \simeq 1160$ sec.

Then the $|M|^2$ nuclear matrix element of the K capture can be written

$$\frac{T_0}{|M|^2} \simeq \frac{\pi}{2} f_K T_K.$$

The $f_K T_K$ product depends only on the β interaction constant (g^2 or T_0) and on the nuclear matrix element of the transition. It is also equal to the $f_+ T_+$ product for β^+ emission, when β^+ emission is in competition with K capture; hence we have in the case of β^+ decay, capture being necessarily present,

$$\frac{f_K T_K}{f_+ T_+} = \frac{2}{\pi}, \quad \text{whence} \quad \frac{T_0}{|M|^2} = \frac{\pi}{2} f_K T \left(1 + \frac{P_+}{P_K}\right). \qquad \ldots\ldots(59)$$

If we want to get $|M|^2$, it is necessary to measure not only the half-life of the decay, but also the branching ratio P_K/P_+. Let us note that the measurement of P_K/P_+ is a good test of the β decay theory, its value being independent of the nuclear properties.

† This new value for the period of the neutron has been reported by Liubimov at the International Congress of Nuclear Physics (Paris 1959). The corresponding value of T_0 is (1170 ± 35) sec.

4.3.3. *Classification of the transitions according to the $f_K T_K$ product.*

The capture transitions can be classified according to Sargent's method (1933) as it has been remarked by Thompson (1949) and developed by Feather (1952 a, b, c) and Major and Biedenharn (1954). But the Sargent diagram (cf. Major and Biedenharn 1954, figure 1) obtained by plotting the probability $1/T_K$ against the energy of the transition, gives a cloud of points from which it is difficult to characterize the degree of forbiddenness of a transition.

It is better to classify the transitions by means of the $f_K T_K$ product, as has been pointed out by Konopinski (1943) for β emission. In the original method of Konopinski, the f_K function to be used corresponds to an allowed transition. The distributions obtained (Major and Biedenharn 1954) make it possible to distinguish quite well between the allowed transitions and the forbidden transitions of different orders.

We should note that Konopinski's method does not classify the transitions according to the values of the matrix elements, since $f_K T_K$ is inversely proportional to $|M|^2$ only for allowed transitions. The classification of the transitions according to the $|M_n|^2$ actual nuclear matrix element can be performed for the capture as well as for the β emission (Bouchez 1950, Nataf and Bouchez 1952); and the nuclear factor R of forbiddenness can be determined by comparing the nuclear matrix element $|M_n|$ calculated from $f_K^{(n)} T_K$ actual product with that $|M_0|$ of the $n \rightarrow p$ allowed transition :

$$R \equiv \frac{|M_n|^2}{|M_0|^2} = \frac{2}{\pi} \frac{T_0}{|M_0|^2} \frac{1}{f_K^{(n)} T_K} \simeq \frac{2}{\pi} \frac{1170}{f_K T_K}. \qquad \ldots\ldots(60)$$

This R factor will be discussed in the study of the allowed and forbidden transitions.

4.4. *Relativistic Eigenfunctions of the Leptons*

It is necessary to take as eigenfunctions for the electron and neutrino the solutions of the Dirac equation describing relativistic particles. On the contrary, the nucleons in the nucleus can be considered as non-relativistic particles (their kinetic energy being about 20 Mev, very much lower than $M_0 c^2 \simeq 930$ Mev) and can be described by Pauli spinors.

It is also convenient to use spherical harmonic functions as angular functions, allowing us to relate easily the transition probability (or more generally the angular distribution and correlation in β emission) with the change in angular momentum and parity between nuclear levels.

4.4.1. *Atomic electron radial functions.*

The relativistic radial functions of the electron in the discrete spectrum are given in first approximation by Bethe (1933) in the case of a point charge and without taking screening into account; they comprise large components $g_K, g_{L_I}, f_{L_{II}}, g_{L_{III}}$ and small components $f_K, f_{L_I}, g_{L_{II}}, f_{L_{III}}$ for the different shells K, L_I, L_{II}, L_{III} ... but the recent measurements on the branching ratios L/K and K/β^+ are accurate enough to require a correction of the above functions for the secondary effects of finite size of the nucleus and screening (these corrections amount to about 10%). These effects were calculated with much refinement by Brysk and Rose (1955, 1958), using machine computation.

Finite size of the nucleus. The potential inside the nucleus ($r < R$) is different from the Coulomb potential ($V = -Ze^2/r$) which exists outside ($r > R$); but the form of the inner potential has not a great influence on the electron radial functions, the only important condition being that it should remain finite at the origin, increase monotonically and join continuously to the Coulomb potential outside the nucleus.

This correction calculated by Rose (1951), Rose and Holmes (1951), Malcolm and Strachan (1951), Malcolm (1952), Nataf (1954, 1956), and Brysk and Rose (1955, 1958), corresponds in fact to a decrease of the Coulomb potential inside the nucleus and evidently tends to decrease the $f(R)$ and $g(R)$ components simultaneously at the nuclear radius $r = R$, the small components being more reduced than the large components.

On the contrary, in the case of β^+ emission, the finite size effect tends to increase the radial functions $g(R)$ and $f(R)$ (Malcolm 1952), which is important for the calculation of the branching ratios K/β^+.

Let us note that the correction for finite size of the nucleus increases with the atomic number Z and is more important for K or L_I functions, because of the proximity of these electrons to the nucleus.

Screening effect. The second correction is due to the screening effect of the other atomic electrons, which is important mostly at a large distance from the nucleus for p electrons. In first approximation the empirical method of Slater (1930) can be applied: hydrogen-like Bethe functions are taken, the atomic number Z and the principal quantum number n being reduced by Slater's screening factor (Marshak 1952).

To get more accurate expressions for the K or L_I radial functions we can use Reitz's (1950) functions calculated with a modified Coulomb potential to take the other atomic electrons into account. But the accuracy of this method is limited by difficulties of interpolation in Reitz's tables. This screening effect was recently evaluated by Brysk and Rose (1955, 1958) using a renormalization method of radial functions; near the nucleus the electron moves in the unscreened nuclear field so that the wave function in this region has the same analytical form as for a 'hydrogen-like' atom, but with an altered normalization factor.

In the case of β^+ emission these corrections are important for soft electrons (modification of the K/β^+ ratio).

Figures 18 to 22 represent the electronic radial functions according to the calculations of Brysk and Rose (1955). These functions take into account the corrections for finite nuclear radius, for screening and for the secondary terms of the series expansion in powers of αZr for the electron, the average being taken over the nuclear volume.

4.4.2. *Neutrino radial functions.*

The neutrino wave functions, solutions of the Dirac equation for mass equal to zero, are analogous to the electron relativistic functions (Rose 1937) when Bessel functions are taken for radial functions. These neutrino functions can be normalized to one particle in a box (Marshak 1942); we shall adopt here the normalization per unit energy interval (Nataf and Bouchez 1952).

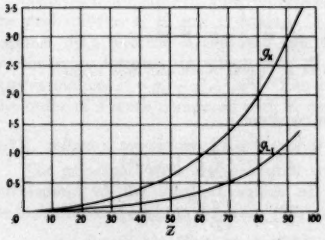


Figure 18. Large components g_K, g_{LI} of the electronic radial functions for K and L_I shells. (*After* Brysk and Rose 1955.)

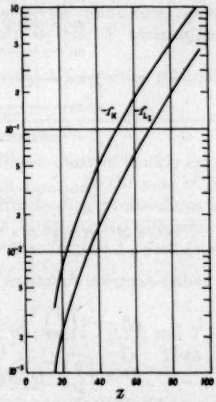


Figure 19. Small components f_K, f_{LI} of the electronic radial functions for K and L_I shells. (*After* Brysk and Rose 1955.)

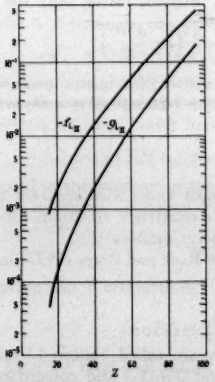


Figure 20. Large component f_{LII} and small component g_{LII} of the electronic radial functions for L_{II} shell. (*After* Brysk and Rose 1955.)

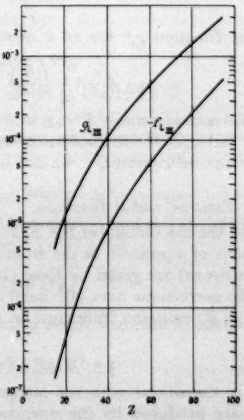


Figure 21. Large component g_{LIII} and small component f_{LIII} of the electronic radial functions for L_{III} shell. (*After* Brysk and Rose 1955.)

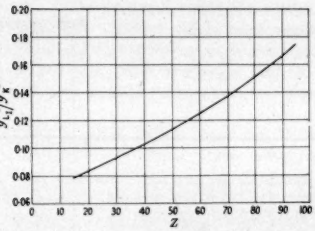


Figure 22. $g_{L_1}^2/g_K^2$ as a function of Z. (*After* Brysk and Rose 1955.)

The radial neutrino functions can be written

$$g_l^B = -\left(\frac{q}{2r}\right)^{1/2} J_{l+1/2}(qr) \qquad g_{-l-1}^B = \left(\frac{q}{2r}\right)^{1/2} J_{l+1/2}(qr)$$

$$f_l^B = \left(\frac{q}{2r}\right)^{1/2} J_{l+3/2}(qr) \qquad f_{-l-1}^B = \left(\frac{q}{2r}\right)^{1/2} J_{l-1/2}(qr).$$

The neutrino energy q being small, in practice less than a few $m_0 c^2$, we can take the approximate expression of the Bessel function for $qR \ll 1$ at the nuclear radius ($r = R$)

$$J_{n+1/2}(qr) = \frac{(2/\pi)^{1/2}(qR)^{n+1/2}}{(2n+1)!!}.$$

So the function g_s^B for an s neutrino can be written, taking only the first two terms,

$$g_s^B = -\left(\frac{q}{2R}\right)^{1/2} J_{1/2}(qR) \simeq -\frac{q}{\sqrt{\pi}}\left(1 - \frac{q^2 R^2}{6}\right)$$

and the second term $q^2 R^2/6 \lesssim 10^{-3}$ (with $q \simeq 1$ and $R \lesssim 0{\cdot}2$ in the usual relativistic units, one uses $R \simeq 0{\cdot}43\alpha A^{1/3}$ corresponding to $R \simeq 1{\cdot}20 \times 10^{-13} A^{1/3}$ cm, α being the fine structure constant); we can omit this second term.

4.4.3. *Positron radial functions.*

For the calculation of the K/β^+ ratio, we must know with accuracy the radial functions of a positron in the atomic field. The relativistic functions for electrons (or positrons) are given by Rose (1937) for a point nucleus.

The corrections have also been calculated by Rose and Perry (1953) in the case of β^- or β^+ emission (influence on the K/β^+ ratio).

§ 5. The Allowed Transitions

The transitions with the largest probability are called ' allowed ' transitions. They are produced by the operators $T_0(Y_0)$ and $T_1(\sigma Y_0)$ and correspond to the selection rule: $\Delta I = 0$ or 1 and $\Delta\pi = 1$ (no parity change). More precisely, when the orbital angular momentum L of the nuclear levels can be defined, the allowed transitions correspond to $\Delta L = 0$. The allowed transitions comprise:

The transitions $\Delta I = 0$ produced either by the V vector interaction (non-relativistic part), operator T_0, or by the A axial-vector interaction, operator T_1; but the $0^{\pm} \rightarrow 0^{\pm}$ transitions are produced only by the V interaction (Fermi transitions).

The transitions $\Delta I = 1$ are produced only by the A interaction (Gamow–Teller transitions).

5.1. *Møller's Formulae*

Generally, for allowed and forbidden transitions, one can distinguish between the probabilities of K and L capture (L_I, L_{II} or L_{III}). In the case of K and L_I capture, an s electron is captured and the corresponding probabilities are given by the same formula, in which only the radial functions are different. In first approximation $P_s \sim g_s^2 \sim \alpha^2 Z^2/n^3$. The contribution for the L_I shell ($n = 2$) is of the order of a tenth compared with K capture ($n = 1$).

5.1.1. *Allowed transitions* $\Delta I = 0$ (*not* $0 \rightarrow 0$).

The principal terms are associated with the operators $T_0(\Upsilon_0)$ and $T_1(\sigma\Upsilon_0)$ and correspond to the smallest values of j_1 and j_2 ($j_1 + j_2 = 1$). Then the total probability of transition is, from (55) and (56),

$$P = 2g^2(2j_2+1)\,[C_V^2\,|\,M_0(\Upsilon_0)|^2\,\|\,L_0\|^2 + \tfrac{1}{3}C_A^2\,|\,M_1(\sigma\Upsilon_0)|^2\,\|\,L_1\|^2]. \quad \ldots\ldots(61)$$

s *capture—Møller's formula.* The K or L_I electron is captured from an s state $|j_1 = \frac{1}{2}, l_1 = 0\rangle$ and a left-handed neutrino is emitted in an (s, p) state with indeterminate parity; the Dirac eigenfunction of this neutrino can be written

$$|j_2\rangle = \frac{1}{\sqrt{2}}\,[|\tfrac{1}{2}, 0\rangle + |\tfrac{1}{2}, 1\rangle].$$

For the V interaction the probability is, from (56),

$$P_s^V = 2g^2\,C_V^2(2j_2+1)\,|\,M_0(\Upsilon_0)|^2\,|\langle \tfrac{1}{2}\,0\,\tfrac{1}{2}|\,T_0^0(\Upsilon_0)\,|\,\tfrac{1}{2}\,0\,\tfrac{1}{2}\rangle|^2$$

the calculation being made for the transition $m_1 = m_2 = \frac{1}{2}$, for which the Clebsch–Gordan coefficient is equal to unity. The lepton term in two-component theory takes the simple form

$$\langle L_0\rangle = \langle\phi|\,T_0^0(\Upsilon_0)\,|\psi\rangle = \tfrac{1}{2}\Upsilon_0^0 g_0^B g_s \quad \text{where} \quad g_0^B = -\,q/\sqrt{\pi}$$

and is equal to the product of the large radial functions g_s and g_0^B of the electron and of the neutrino. Whence Møller's formula (1937 a, b)

$$P_s^V = \frac{g^2}{4\pi^2}\,C_V^2\,|\,M_0(\Upsilon_0)|^2\,q^2 g_s^2 \qquad \ldots\ldots(62a)$$

for s capture and V interaction. We get a similar result with the A interaction

$$P_s^A = \frac{g^2}{4\pi^2}\,C_A^2\,|\,M_1(\sigma\Upsilon_0)|^2\,q^2 g_s^2 \qquad \ldots\ldots(62b)$$

and we have the total probability

$$P_s = \frac{g^2}{4\pi^2}\,|\,M|^2\,q^2 g_s^2 \qquad \ldots\ldots(62c)$$

where

$$|\,M|^2 \equiv C_V^2\,|\,M_0(\Upsilon_0)|^2 + C_A^2\,|\,M_1(\sigma\Upsilon_0)|^2.$$

L_{II} *capture.* If a $p(L_{II})$ electron is captured, the neutrino emitted in the state $j_\beta = \frac{1}{2}$ gives the largest contribution to the probability; so we get the same expression as for capture of an s electron

$$P_{L_{II}} = \frac{g^2}{4\pi^2} |M|^2 q^2 f_{L_{II}}{}^2 \qquad \ldots\ldots(63)$$

but with $f_{L_{II}}$ for the radial function of the L_{II} electron.

L_{III} *capture.* The L_{III} capture for allowed transitions is always completely negligible. The electron is captured in a p state ($j_1 = \frac{3}{2}$); from the leading term $\langle \frac{1}{2} | T_1 | \frac{3}{2} \rangle$ we see that

$$P_{L_{III}} = \frac{g^2}{4\pi^2} C_A{}^2 |M_1|^2 q^4 R^2 g_{L_{III}}{}^2 \qquad \ldots\ldots(64)$$

is very small $[g^2(L_{III})/f^2(L_{II}) \simeq R^2 \simeq 10^{-4}]$. And the V interaction gives a smaller term $\langle \frac{3}{2} | T_0 | \frac{3}{2} \rangle$, the neutrino being emitted in a $\frac{3}{2}$ state.

5.1.2. *Allowed transitions* $0^\pm \to 0^\pm$.

Only the operator $T_0(Y_0)$ of the V vector interaction can produce this Fermi transition. The probability is given by Møller's formula (62*a*), the nuclear matrix element $|M_0|^2$ being about four times smaller than that of a normal transition $\Delta I = 0$ ((V, A) interaction).

5.1.3. *Allowed transitions* $\Delta I = 1$.

For the Gamow–Teller transitions with $\Delta I = 1$ only the operator $T_1(\sigma Y_0)$ of the A interaction can operate. The probability is still given by Møller's formula, but with the nuclear matrix element $|M_1(\sigma Y_0)|^2$ a little smaller than that of a normal transition $\Delta I = 0$.

5.2. *Experimental Results*

Super-allowed transitions: type ^{7}Be.

If we consider the distribution of the fT values for the transitions by capture, it can be seen that there is a first important group between 2×10^4 and 2×10^5 seconds, corresponding to the allowed transitions. But before this group there exist some isolated transitions with an fT product which is smaller, its order of magnitude being 10^3 seconds. This small group is analogous to the 'super-allowed' transitions between exceptionally similar nuclei (for instance, the mirror nuclei).

A typical 'super-allowed' transition is that of the ^{7}Be to its mirror nuclide ^{7}Li. It is well known that, in fact, there are two partial transitions (figure 1), the first to the ground state $P_{3/2}$ and the second to the excited state $P_{1/2}$ of ^{7}Li:

$$P_{3/2} \xrightarrow{K_0} P_{3/2} \quad 88\% \quad T_0/|M_0|^2 = \tfrac{1}{2}\pi f_K{}^0 T_K{}^0 = 1900 \text{ sec}$$

$$P_{3/2} \xrightarrow{K_1} P_{1/2} \quad 12\% \quad T_0/|M_1|^2 = \tfrac{1}{2}\pi f_K{}^1 T_K{}^1 = 2700 \text{ sec}$$

the functions $f_K \equiv q_K{}^2 g_K{}^2$ corresponding to allowed transitions. The values of the parameters $T_0/|M|^2$ are hardly larger than the neutron value. This indicates that the transitions take place between nuclear states which are very much alike.

Normal allowed transitions.

The normal allowed transitions, obeying the selection rule ($\Delta I = 0$ or 1, $\Delta\pi = +1$) are characterized by an fT value of the order 10^5 seconds. Let us quote as examples :

$1d_{3/2}$ *shell.* The pure transition

$$^{37}A\,(\tfrac{3}{2}^+) \rightarrow Cl\,(\tfrac{3}{2}^+)$$

corresponds to the transformation of a proton into a neutron inside the $d_{3/2}$ shell. The fT product is about 10^5 sec.

$1f_{7/2}$ *shell.* The transition (K, β^+)

$$^{45}Ti\,(\tfrac{7}{2}^-) \rightarrow Sc\,(\tfrac{7}{2}^-)$$

corresponds to the transformation $p(f_{7/2}) \rightarrow n(f_{7/2})$ and the fT product is 4×10^4 sec.

2p *shell.* The transition (K, β^+)

$$^{61}Cu\,(\tfrac{3}{2}^-) \rightarrow Ni\,(\tfrac{3}{2}^-)$$

corresponds to a transition in the same shell $p(p_{3/2}) \rightarrow n(p_{3/2})$ with $fT \simeq 8{\cdot}5 \times 10^4$ sec.

The pure transition

$$^{71}Ge\,(\tfrac{1}{2}^-) \rightarrow Ga\,(\tfrac{3}{2}^-)$$

corresponds to the transformation $p(p_{3/2}) \rightarrow n(p_{1/2})$ with $fT \simeq 2{\cdot}3 \times 10^4$ sec.

2d *shell.* One might think of a $2d_{5/2}$ proton transforming in a $2d_{3/2}$ neutron leading to a rather small nuclear element, but the pure transition

$$^{131}Cs\,(\tfrac{5}{2}^+) \rightarrow Ba\,(\tfrac{3}{2}^+)$$

has $fT \simeq 3 \times 10^5$ sec. This shows a relatively weak decrease of the nuclear matrix element, if compared with the transition $2d_{3/2} \rightarrow 2d_{3/2}$ for which $fT \simeq 10^5$ sec.

Unfavoured allowed transitions.

The super-allowed transitions being allowed transitions which are favoured by nuclear structure, it seems reasonable to expect allowed transitions ($\Delta I = 0$ or 1, $\Delta\pi = +1$) but unfavoured by nuclear structure. Let us quote two types of such an interaction.

(*a*) *Transitions unfavoured by orbital momentum.* An example is the transition (K, β^+)

$$^{65}Zn\,(\tfrac{5}{2}^-) \rightarrow Cu\,(\tfrac{3}{2}^-)$$

with $fT \simeq 2{\cdot}5 \times 10^7$ sec which indicates a forbidden transition in spite of $\Delta I = 1$ and $\Delta\pi = +1$. The study of this transition (Bouchez *et al.* 1953) and particularly of the K/β^+ branching ratio shows that the nuclear matrix element is considerably decreased by the weak proportion of the actually allowed transitions. If the one-particle model is considered, the decay of ^{65}Zn corresponds to an *l*-forbidden transition ($\Delta L = 2$)

$$p(f_{5/2}) \rightarrow n(p_{3/2})$$

which can be produced only by $T_2(\sigma Y_2)$ or $T_2(Y_2)$, these operators giving a probability about 10^5 times smaller than it would be with the operators $T_0(Y_0)$ and $T_1(\sigma Y_0)$. This one-particle model is certainly not adequate in this case and the transition is not merely $F_{5/2} \rightarrow P_{3/2}$ but is more complex, with a matrix element of the form

$$\langle aP_{3/2} + bF_{5/2} | H_\beta | F_{5/2} \rangle$$

with $|a|^2+|b|^2=1$ and $|b|^2 \ll |a|^2$. The result is that, as long as $|b|^2 \ll 1$ but is greater than 10^{-5}, the transition probability is given by

$$P \simeq |b|^2 |\langle F_{5/2}| T_1(\sigma Y_0)| F_{5/2}\rangle|^2.$$

All the characteristics of the transition (shape of the β^+ spectrum, branching ratio K/β^+) belong to the allowed transitions (operator T_1) but the fT product is lowered by the very small value of $|b|^2$. In this case we cannot speak of an l-forbidden transition, but of an allowed transition which is unfavoured by the weak proportion of $\Delta L = 0$ transitions.

(*b*) *Reduction of transition probability by nuclear structure change.* When the proton → neutron transition is accompanied by a drastic change of nuclear structure, the nuclear matrix element is reduced and the fT product is increased. Let us take as an example of nuclear structure unfavoured transition the pure transition

$$^{55}\mathrm{Fe}(\tfrac{3}{2}^-) \rightarrow \mathrm{Mn}(\tfrac{5}{2}^-)$$

where the fT product is 9×10^5 sec. The transition concerns more than one nucleon. In fact, according to the one-particle model, an $f_{7/2}$ proton should transform into a neutron on the $p_{3/2}$ shell, with $\Delta I = 2$. There is also a rearrangement of at least three nucleons of the $f_{7/2}$ shell, in order to give a final state $(\frac{5}{2}^-)$. This nuclear rearrangement reduces the overlap of the nuclear eigenfunctions and also the nuclear matrix element.

Nuclear reduction factor.

Introducing (see equation (60)) the nuclear reduction factor R, the preceding discussion can be summed up as follows :

Super-allowed transitions (type ^{7}Be)	$R \simeq 0{\cdot}5$
Normal-allowed transitions (type ^{37}A)	$0{\cdot}1 \leqslant R \leqslant 10^{-2}$
Unfavoured allowed transitions (type ^{55}Fe)	$R \leqslant 10^{-3}$

Let us note that the R factor is a magnitude which is very helpful when studying nuclear structure and the coupling schemes between nucleons ; these reduction factors come into play not only in electron capture and β decay, but also in γ emission and α radioactivity. The analysis of these different nuclear factors has not yet been achieved.

§ 6. Forbidden Transitions with 'Unique' Matrix Element

6.1. *Theoretical Formulae*

In β emission, some transitions are characterized by a unique nuclear matrix element. They correspond to the β spectra with an α shape ($\Delta I = 2$, $\Delta\pi = -1$) or D_2 shape ($\Delta I = 3$, $\Delta\pi = +1$). These transitions are associated with the unique operator $T_{n+1}(\sigma Y_n)$ and correspond to $\Delta I = n+1$ and $\Delta\pi = (-1)^n$, or more simply $\Delta L = n$ if ΔL can be defined. The above analysed allowed transitions ($\Delta I = 1$, $\Delta\pi = 1$) belong to the same class and are produced by the operator $T_1(\sigma Y_0)$. Let us note that these transitions are produced by the A interaction only and evidently cannot give any indication about the mixing of (A, V) interaction ; but they are a very good test of the theory of β decay, since the shape of the β spectra

is very characteristic (α or D_s shape) and also the value of the K/β^+ ratio is independent of the $|M|^2$ nuclear matrix element.

s *capture.*

The probability of transition is given by the general relation

$$P_s = 2g^2 C_A^2 |M_{n+1}(\sigma Y_n)|^2 |\langle n+\tfrac{1}{2} \| T_{n+1}(\sigma Y_n) \| \tfrac{1}{2} \rangle|^2$$

the neutrino being emitted into the state $j_s = n+\frac{1}{2}$; the calculation gives

$$P_s = \frac{g^2}{4\pi^2} C_A^2 |M_{n+1}(\sigma Y_n)|^2 \frac{(qR)^{2n}}{[(2n+1)!!]^2} q^2 g_s^2. \qquad \ldots\ldots(65)$$

For $n = 0$ we get Møller's formula for the allowed transition $\Delta I = 1$.

p *capture* (L_{II} *or* L_{III}).

The lepton matrix elements can be written

L_{II} capture: $\langle n+\frac{1}{2} \| T_{n+1}(\sigma Y_n) \| \frac{1}{2} \rangle$

L_{III} capture: $\langle n-\frac{1}{2} \| T_{n+1}(\sigma Y_n) \| \frac{3}{2} \rangle$

and lead to the probabilities

$$P_{L_{II}} = \frac{g^2}{4\pi^2} C_A^2 |M_{n+1}(\sigma Y_n)|^2 \frac{(qR)^{2n}}{[(2n+1)!!]^2} q^2 f_{L_{II}}^2$$

$$P_{L_{III}} = \frac{g^2}{4\pi^2} C_A^2 \frac{3n}{2n+1} |M_{n+1}(\sigma Y_n)|^2 \frac{(qR)^{2n-2}}{[(2n-1)!!]^2} q^2 g_{L_{III}}^2. \qquad \ldots\ldots(66a)$$

Let us note that L_{III} capture is comparable with L_{II} capture

$$\frac{L_{III}}{L_{II}} \simeq 3n(2n+1) \frac{1}{q^2 R^2} \frac{g_{L_{III}}^2}{f_{L_{II}}^2} \qquad \ldots\ldots(66b)$$

with $$g_{L_{III}}^2 / f_{L_{II}}^2 \simeq R^2.$$

6.2. *Typical Experimental Transition*: ^{122}Sb (2·73d)

The odd–odd nuclide ^{122}Sb decays either by β^- emission or by β^+ emission and electron capture. The β^- spectrum (28%) for transition to the ground state (0^+) of ^{122}Te has the α shape, which characterizes a transition ($\Delta I = 2$ and $\Delta\pi = -1$) with unique matrix element. The ground state of ^{122}Sb is therefore 2^-. It follows that the transition by capture towards the even–even ground state of ^{122}Sn

$$^{122}\text{Sb}\,(2^-) \rightarrow \text{Sn}\,(0^+)$$

is a ' unique ' forbidden transition. The experimental investigation of this nuclide has been carried out by Glaubmann (1955), Farrelly *et al.* (1955) and recently by Perlman *et al.* (1958). The ratio of the K_0 transition is about 2·1% and the maximum kinetic energy of the positrons about (565 ± 25) kev. The K capture probability can be written according to (65)

$$P_K = \frac{g^2}{4\pi^2} C_A^2 |M_2|^2 \frac{q^2 R^2}{9} q^2 g_K^2 \qquad \ldots\ldots(67)$$

which permits the calculation of the fT product for the actual K_0 transition

$$T_0/C_A^2|M_2|^2 = \tfrac{1}{2}\pi f_K T_K \simeq 1{\cdot}7 \times 10^4 \text{ sec}$$

with $q \simeq 3{\cdot}1 m_0 c^2$. This value is about 10^2 times higher than that of a super-allowed transition whence the nuclear R factor of forbiddenness is only from (60),

$$R \equiv |M_2|^2/|M_0|^2 \simeq 6 \times 10^{-2}.$$

The comparison with the R nuclear factor corresponding to the β_0^- emission to the ground state is very interesting. We obtain with $\beta_0^- = 28\%$, $\epsilon_0 = 4{\cdot}90\, m_0 c^2$,

$$T_0/|M_0^-|^2 = T_- \frac{1+\gamma_0}{2} \int_1^{\epsilon_0} \epsilon p q^2 F_0 \, d\epsilon = 4 \times 10^3 \text{ sec}$$

$$T_0/|M|^2 = T_- \frac{1+\gamma_0}{2} \frac{R^2}{9} \int_1^{\epsilon_0} \epsilon p q^2 F_0 \left(q^2 + 9\frac{L_1}{L_0}\right) \simeq 1{\cdot}18 \times 10^5 \text{ sec}$$

using Greuling's notation (cf. § 10.2.1). The nuclear forbiddenness factor for β_0^- emission is therefore 8×10^{-3}, that is to say about seven times smaller than that for K capture. The transition K_0 or β_0^+ appears then as favoured when compared with β^- emission. Glaubmann (1955) using the approximate rule $\log[fT(\epsilon_0^2 - 1)] \simeq 10$ for transitions of the type ($\Delta I = 2, \Delta\pi = -1$), obtained about ten for the ratio of fT products for the two transitions. It seems there are two reasons for the $p \to n$ transition being more favoured than the $n \to p$ one. On the one hand, the fundamental final states have not the same configurations ; the magic nucleus $^{122}_{50}$Sn should be practically pure $^1S_0$ but $^{122}_{52}$Te can be a mixture ($^1S_0 + {}^5D_0$) giving rise to a relative reduction of the β^- transition rate. On the other hand, the unoccupied states of the $h_{11/2}$ shell are more numerous than those of the $g_{7/2}$ shell, increasing the transition probability of a proton towards an $h_{11/2}$ neutron state.

§ 7. Parity-forbidden Transitions

The transitions with $\Delta I = 0$ or 1, $\Delta\pi = -1$ are parity-forbidden. Their probability is one order of magnitude smaller than that of the allowed transitions. These first-order forbidden transitions are produced by operators of rank 0 or 1 with odd parity, and fall into three groups :

(*a*) the transitions $\Delta I = 1$ connected with the operators of rank 1 and 2, $T_1(Y_1)$, $T_1(\sigma Y_1)$, $T_1(\alpha Y_0)$ and $T_2(\sigma Y_1)$. The principal lepton terms correspond to $j_1 + j_2 = 1$ ($j_1 = j_2 = \frac{1}{2}$) and $j_1 + j_2 = 2$ ($j_1 = \frac{1}{2}$, $j_2 = \frac{3}{2}$ and the reverse). For instance, in s capture ($j_1 = \frac{1}{2}$) the neutrino is emitted in a state $j_2 = \frac{1}{2}$ or $\frac{3}{2}$.

(*b*) the transitions $\Delta I = 0$ (except $0^+ \to 0^-$) connected not only with the above operators of rank 1 but also with the odd operators of rank 0 : $T_0(\sigma Y_1)$ and $T_0(\gamma^5 Y_0)$, the neutrino being emitted only in the state $j_2 = \frac{1}{2}$ in the case of s or L_{II} capture. On the contrary these operators of rank zero do not come into play for L_{III} capture.

(*c*) the transitions $0^+ \leftrightarrow 0^-$ connected only with the pseudoscalar operators $T_0(\sigma Y_1)$ and $T_0(\gamma^5 Y_0)$. L_{III} capture is forbidden for these transitions.

7.1. *Parity-forbidden Transitions* $\Delta I = 1$

The operators $T_1(Y_1)$, $T_1(\sigma Y_1)$ and $T_1(\alpha Y_0)$, irreducible tensors of rank 1, and the $T_2(\sigma Y_1)$ tensor contribute to this type of transition

V interaction
$$\begin{cases} C_V^2 |M_1(Y_1)|^2 \sum_{j_2} (2j_2+1) \| L_1(Y_1) \|^2 \\ C_V^2 |M_1(\alpha Y_0)|^2 (2j_2+1) \| L_1(\alpha Y_0) \|^2 \\ 2C_V^2 M_1(Y_1) M_1(\alpha Y_0) (2j_2+1) \| L_1(Y_1) \| . \| L_1(\alpha Y_0) \|^* \end{cases} \quad \ldots\ldots(68a)$$

A interaction
$$\begin{cases} C_A^2 |M_1(\sigma Y_1)|^2 \sum_{j_2} (2j_2+1) \| L_1(\sigma Y_1) \|^2 \\ C_A^2 |M_2(\sigma Y_1)|^2 (2j_2+1) \| L_2(\sigma Y_1) \|^2 \end{cases} \quad \ldots\ldots(68b)$$

(V, A) interaction
$$\begin{cases} 2C_A C_V M_1(Y_1) M_1(\sigma Y_1) \sum_{j_2} (2j_2+1) \| L_1(Y_1) \| . \| L_1(\sigma Y_1) \|^* \\ 2C_A C_V M_1(Y_1) M_1(\alpha Y_0) (2j_2+1) \| L_1(\alpha Y_0) \| \| L_1(\sigma Y_1) \|^* . \end{cases} \quad \ldots\ldots(68c)$$

There first appear the square terms corresponding to the matrix elements $\langle j_2 | T_1 | j_1 \rangle$ and $\langle j_2 | T_2 | j_1 \rangle$ calculated for $j_1 + j_2 = 1$ and $j_1 + j_2 = 2$. For the operator $T_1(\alpha Y_0)$ the term $j_1 + j_2 = 2$ contains neutrino functions with $j_2 = \frac{3}{2}$ and gives a negligible contribution. For the operator $T_2(\sigma Y_1)$ the neutrino is forced towards $j_2 = \frac{3}{2}$ by the selection rule if $j_1 = \frac{1}{2}$. There are also cross terms arising from the interactions between the tensors T_1. These interaction terms do not vanish, because the ratios of the different matrix elements are real.

Finally, the probabilities for the different captures can be derived from tables 4 and 5 with $n = 1$.

7.2. *Parity-forbidden Transitions* $\Delta I = 0$

When $\Delta I = 0$ with parity change, one must add to the above terms the contribution of the operators of rank zero, $T_0(\alpha Y_1)$ and $T_0(\gamma^5 Y_0)$. This contribution contains only A interaction terms.

$$\left.\begin{array}{l} C_A^2 |M_0(\gamma^5 Y_0)|^2 (2j_2+1) \| L_0(\gamma^5 Y_0) \|^2 \quad \text{for} \quad T_0(\gamma^5 Y_0) \\ C_A^2 |M_0(\sigma Y_1)|^2 (2j_2+1) \| L_0(\sigma Y_1) \|^2 \quad \text{for} \quad T_0(\sigma Y_1) \\ 2C_A^2 M_0(\gamma^5 Y_0) M_0(\sigma Y_1) (2j_2+1) \| L_0(\sigma Y_1) \| . \| L_0(\gamma^5 Y_0) \|^* \\ \qquad \text{for the interference of } \gamma^5 Y_0 \text{ and } \sigma Y_1. \end{array}\right\} \quad \ldots\ldots(69)$$

The summation over j_2 is useless in the case of the operator $T_0(\gamma^5 Y_0)$ and also for the cross term. Only the case $j_1 + j_2 = 1$ is of importance: the terms with $j_1 + j_2 = 2$ are negligible. We obtain for K, L_I and L_{II} captures:

$$P^A = \frac{g^2}{4\pi^2} C_A^2 q^2 \left[|M_0(\gamma^5 Y_0)|^2 g^2 + |M_0(\sigma Y_1)|^2 \left(f^2 + 2fg \frac{qR}{3} + g^2 \frac{q^2 R^2}{9} \right) + 2M_0(\gamma^5 Y_0)^* M_0(\sigma Y_1) g \left(f + g \frac{qR}{3} \right) \right] \quad \ldots\ldots(70)$$

for P_K one puts $g = g_K$, $f = f_K$

for P_{L_I} one puts $g = g_{L_I}$, $f = f_{L_I}$

for $P_{L_{II}}$ one puts $g = f_{L_{II}}$, $f = -g_{L_{II}}$.

The probability is practically negligible for L_{III} capture.

7.3. *Typical Experimental Transition* : ^{235}Np

The recent experimental results of Gindler *et al.* (1958) on the decay scheme of ^{235}Np, and also the spin measurements on ^{235}U ($\frac{7}{2}$) and ^{237}Np ($\frac{5}{2}$) together with the shell model indicate that the transition $^{235}\mathrm{Np}(\frac{5}{2}^-) \rightarrow {}^{235}\mathrm{U}(\frac{7}{2}^+)$ is $\Delta I = 1$ and parity-forbidden. The available energy $\Delta M_a = 123$ kev in the K_0 transition has been computed by Gindler *et al.* (1958) from the experimental result L/K = 36·7 and assuming an allowed transition. If we consider the transition as forbidden one gets a value of ΔM_a which is very close to the value of an allowed transition. The fT product of the transition K_0 considered as allowed gives

$$T_0/|M_0^K|^2 \simeq 4{\cdot}75 \times 10^6 \text{ sec}$$

with $\Delta M_a = 126$ kev (cf. § 9) ; this is a value which corresponds to a first-forbidden transition. The exact value for the actual transition is

$$T_0/|M_1|^2 = \tfrac{1}{2}\pi f_1 T \simeq 2 \times 10^6 \text{ sec}$$

taking $|M_1| \simeq C_A |M(\sigma Y_1)| \simeq C_V |M(Y_1)|$ and $|M(\alpha Y_0)| \simeq (\bar{v}/c)|M(Y_1)|$,

$\bar{v}$ being the mean velocity of the nucleon which undergoes the transition.

It results in a nuclear reduction factor $R \simeq 5 \times 10^{-4}$, much smaller than that generally obtained. But it is well known that for heavy nuclei the number of configurations is very large and this causes a reduction of the covering of the nuclear initial and final eigenfunctions.

§ 8. Forbidden Transitions of nth Order

The forbidden transitions of nth order $\Delta I = n$, $\Delta\pi = (-1)^n$ have several matrix elements and are produced by the irreducible tensors of nth rank $T_n(Y_n)$, $T_n(\sigma Y_n)$, $T_n(\alpha Y_{n-1})$ and of $(n+1)$th rank $T_{n+1}(\sigma Y_n)$, the parity-forbidden transitions ($\Delta I = 1$) being a special case corresponding to $n = 1$.

The final formulae are given in table 4 for the K, L_I and L_{II} shells, and in table 5 for the L_{III} shell.

§ 9. L Capture and K Capture

Since the last review article by Robinson and Fink (1955), progress has been made in the determination of the relative probabilities of K capture and L capture, and concerns :

(*a*) methods of measurement with proportional counters, scintillation spectrometers, and the associated electronic devices (selection of amplitude or of time) (cf. § 1.3), (*b*) the knowledge of the K and L shell fluorescence yields (cf. § 1.3.2), (*c*) the determination of electron densities at the nuclear surface, the secondary effects of screening by atomic electrons and of the more exact estimation of the finite size of the nucleus (cf. § 4.4) (Rose and Perry 1953, Brysk and Rose 1955, 1958).

9.1. *Theoretical Results for L/K Ratio*

As a result of the theoretical forecast (cf. § 1.1) of the L capture by Marshak (1942), it is well known that the value of the L/K ratio increases when the energy of the transition decreases (^{235}Np, ^{103}Pd, ^{40}K) or when the atomic number increases. The precise formulae giving the L/K ratio can be obtained from the theoretical calculations of the preceding section.

Table 4. Formulae for Orbital Capture Transition Probabilities for K, L_I and L_{II} Shells (Interaction V, A)

$$P^{V} = \frac{g^2}{4\pi^2} C_V^2 q^2 \frac{(qR)^{2n-2}}{[(2n-1)!!]^2} \left\{ \frac{n}{2n+1} |M_n(Y_n)|^2 \left[f^2 + 2fg \frac{qR}{2n+1} + g^2 \frac{q^2R^2}{n(2n+1)} \right] + |M_n(\alpha Y_{n-1})|^2 g^2 - 2\sqrt{\left(\frac{n}{2n+1}\right)} M_n(Y_n)^* M_n(\alpha Y_{n-1}) g\left(f + g\frac{qR}{2n+1}\right) \right\}$$

$$P^{A} = \frac{g^2}{4\pi^2} C_A^2 q^2 \frac{(qR)^{2n-2}}{[(2n-1)!!]^2} \left\{ \frac{n+1}{2n+1} |M_n(\sigma Y_n)|^2 \left[f^2 - 2fg \frac{qR}{2n+1} + g^2 \frac{q^2R^2}{(n+1)(2n+1)} \right] + |M_{n+1}(\sigma Y_n)|^2 g^2 \frac{q^2R^2}{(2n+1)^2} \right\}$$

$$P^{VA} = \frac{g^2}{4\pi^2} C_V C_A q^2 \frac{(qR)^{2n-2}}{[(2n-1)!!]^2} \left\{ 2\sqrt{\left(\frac{n+1}{2n+1}\right)} M_n(\sigma Y_n)^* M_n(\alpha Y_{n-1}) g\left(f - g\frac{qR}{2n+1}\right) - \frac{2\sqrt{[n(n+1)]}}{2n+1} M_n(Y_n)^* M_n(\sigma Y_n) f^2 \right\}$$

to get P_K one puts $f = f_K$, $g = g_K$

to get P_{L_I} one puts $f = f_{L_I}$, $g = g_{L_I}$

to get $P_{L_{II}}$ one puts $f = g_{L_{II}}$, $g = -f_{L_{II}}$

Table 5. Formulae for Orbital Capture Transition Probabilities for L_{III} Shell (Interaction V, A)

$$P_{L_{III}}{}^{V} = \frac{g^2}{4\pi^2} C_V^2 q^2 \frac{(qR)^{2n-4}}{[(2n-3)!!]^2} \frac{3(n-1)}{2n-1} \left\{ \frac{n}{2n+1} |M_n(Y_n)|^2 \left[f^2 + 2fg \frac{qR}{2n-1} + g^2 \frac{2q^2R^2}{3(n-1)(2n-1)} \right] + |M_n(\alpha Y_{n-1})|^2 g^2 - 2\sqrt{\left(\frac{n}{2n+1}\right)} M_n(Y_n)^* M_n(\alpha Y_{n-1}) g\left(f + g\frac{qR}{2n-1}\right) \right\}$$

$$P_{L_{III}}{}^{A} = \frac{g^2}{4\pi^2} C_A^2 q^2 \frac{(qR)^{2n-4}}{[(2n-3)!!]^2} \frac{3(n-1)}{2n-1} \left\{ \frac{n+1}{2n+1} |M_n(\sigma Y_n)|^2 \left[f^2 - 2fg \frac{qR}{2n-1} + g^2 \frac{q^2R^2}{3(n^2-1)} \right] + \frac{n}{n-1} \frac{2n-1}{2n+1} |M_{n+1}(\sigma Y_n)|^2 g^2 \frac{q^2R^2}{(2n-1)^2} \right\}$$

$$P_{L_{III}}{}^{VA} = \frac{g^2}{4\pi^2} C_V C_A q^2 \frac{(qR)^{2n-4}}{[(2n-3)!!]^2} \frac{3(n-1)}{2n-1} \left\{ 2\sqrt{\left(\frac{n+1}{2n+1}\right)} M_n(\sigma Y_n)^* M_n(\alpha Y_{n-1}) g\left(f - g\frac{qR}{2n-1}\right) - \frac{2\sqrt{[n(n+1)]}}{2n+1} M_n(Y_n)^* M_n(\sigma Y_n) \left[f^2 - g^2 \frac{q^2R^2}{3(n-1)(2n-1)} \right] \right\}$$

with $f = f_{L_{III}}$, $g = g_{L_{III}}$

9.1.1. *Allowed transitions.*

The main contribution in L capture comes from L_I capture, but for large values of Z, L_{II} capture cannot be neglected. According to Møller's formulae (60)–(62)

$$\frac{L_I + L_{II}}{K} = \frac{q_{L_I}^2 g_{L_I}^2 + g_{L_{II}}^2 f_{L_{II}}^2}{q_K^2 g_K^2} \simeq \frac{q_{L_{II}}^2 g_{L_I}^2}{q_K^2 g_K^2}\left(1 + \frac{f_{L_{II}}^2}{g_{L_I}^2}\right); \qquad \ldots\ldots(71)$$

this approximate expression is obtained if the energy difference between the L_I and L_{II} absorption discontinuities is neglected. It can be seen that the L_{II}/K ratio, coming into play through the correction term $f_{L_{II}}^2/g_{L_I}^2$ amounts to about 10% (figures 18 and 20) of the principal term L_I/K (figure 22) for intermediate Z values, and cannot be neglected for large values of Z.

If the energy of the transition is large compared with the discontinuities of absorption, i.e. $\Delta M_a \gg E_K$, the relation (71) can be still more simplified to

$$\frac{L}{K} \simeq \frac{g_{L_I}^2}{g_K^2}\left(1 + \frac{f_{L_{II}}^2}{g_{L_I}^2}\right) \qquad \ldots\ldots(72)$$

and is merely equal to the ratio (figure 22) of the electron densities K and L_I in nuclear volume (except for the corrective factor due to the L_{II} electrons).

9.1.2. *Forbidden 'unique' transitions* $\Delta I = n+1$, $\Delta\pi = (-1)^n$.

The relations (65) and (66) for K, L_I and L_{II} capture enable us to write

$$\frac{L_I + L_{II}}{K} = \frac{q_{L_I}^{2n+2} g_{L_I}^2 + q_{L_{II}}^{2n+2} f_{L_{II}}^2}{q_K^{2n+2} g_K^2} \simeq \left(\frac{q_{L_I}}{q_K}\right)^{2n+2} \frac{g_{L_I}^2}{g_K^2}\left(1 + \frac{f_{L_{II}}^2}{g_{L_I}^2}\right). \qquad \ldots\ldots(73)$$

This L/K ratio is equal to that of the allowed transitions if $q_{L_I} \simeq q_K$ but becomes larger if the energy of the transition is low ($\Delta M_a \geqslant E_K$ and $q_{L_I} > q_K$).

The L_{III} capture must be considered in this transition type; we have seen (66*b*) that $L_{III}/L_{II} \simeq 3n(2n+1)/q^2$ indicating for the L_{III} shell a contribution sometimes more important than that of the L_{II} or L_I shells.

9.1.3. *Parity-forbidden transitions.*

The precise formulae are rather lengthy because of the presence of several nuclear matrix elements, but the L and K probabilities are bilinear combinations of the radial functions f and g, with the same coefficients in both cases, and the L/K ratio is very close to that of the allowed transitions. We shall discuss these transitions for ^{235}Np, ^{126}I and ^{113}Sn.

9.2. *Experimental Results for L/K Ratio—Comparison with Theory*

9.2.1. *Allowed transitions.*

^{7}Be

The variation of half-life between the metallic state and the fluoride is due to L capture, but the measured value $\Delta T/T = 10^{-3}$ ($\pm 10\%$) (Leininger *et al.* 1949, Kraushaar *et al.* 1943, Bouchez *et al.* 1956) cannot easily be related to the L/K ratio because the proportion of 2s electrons in the metal and the fluoride is not well known. The calculated value from the formula (72), using Hartree wave functions, is $L/K \simeq 10^{-2}$ (Benoist 1949) but it is not justified to compare this with the experimental value.

^{37}A

It was ^{37}A on which Pontecorvo *et al.* (1949) demonstrated L capture, by means of a proportional counter: the relative intensities of the L and K lines of ^{37}Cl, obtained with a xenon counter (the escape of Cl K_α radiation being negligible) gave directly $L/K \simeq 8{\cdot}5\%$. This experimental result has recently been obtained more precisely by Langevin (1955 c), using a similar method, as $L/K = (9{\cdot}2^{+1}_{-0{\cdot}5})\%$. The value $L/K = 8{\cdot}2\%$ (Benoist 1949) calculated using Hartree functions is in good agreement with the experimental value.

However there remains to be explained the very small number of Auger electrons accompanying K capture. According to the K fluorescence yield (7·5%), about 84% of the transitions should be accompanied by Auger electrons (2400 ev), but Kofoed-Hansen (1954) find only 65%.

^{71}Ge

The direct measurement (Drever and Moljk 1957) of the L/K ratio in ^{71}Ge with the proportional anticoincidence counter described in § 1.4, avoiding the corrections due to Ga K_α escape, leads to $L/K = 0{\cdot}128$ ($\pm 4\%$). This result is also in agreement with the older measurements (Langevin 1955 a, Laberrigue-Frolow and Radvanyi 1956) made with an ordinary proportional counter and depending on R_K(Ga): L/K varies from 0·11 to 0·19 if R_K is taken between 0·55 and 0·50 (the probable value for R_K is 0·50, from figure 13, and corresponds well to Drever and Moljk's (1957) measurement: $0{\cdot}53 \pm 0{\cdot}03$).

The theoretical calculation gives $L/K = 0{\cdot}12$ taking $\Delta M_a = 233$ kev (Langevin 1955 a, Bisi 1955) obtained from γ-spectra of internal bremsstrahlung and using the radial functions given by Brysk and Rose (1955). The L_{II} capture, which has been taken into account, gives a small contribution, only 1%. The agreement is thus excellent.

^{79}Kr

The measurement carried out by Radvanyi (1954, 1956) with an ordinary proportional counter makes allowance for a correction due to the escape of part of the K_α radiation and depends on the fluorescence yield R_K of bromine. If one takes for R_K(Br) a value between 0·60 and 0·62, Radvanyi's result is $13\% < L/K < 18\%$ for all the transitions by capture. The L/K value for the transition to the ground state is probably a little smaller, but this correction cannot exceed 10%, the transitions to the excited states of ^{79}Br being at most 30%.

The calculation gives $L/K \simeq 10{\cdot}9\%$ using the functions of Brysk. This value is smaller than the experimental value. But before stating that this difference is significant, it is necessary to make a direct measurement of L/K, independently of R_K.

^{103}Pd

There is no precise measurement of the L/K ratio for ^{103}Pd (K_1 and L_1 transitions to the first excited state of ^{103}Rh, half-life 57 minutes). The experiments (Flack and Mason 1958, Saraf 1954, Bouchez *et al.* 1955) indicate a value of L/K between 5 and 10%.

The theoretical value, given merely by $g_{L_I}^2/g_K^2$, is about 11%.

^{109}Cd

The transition to the 40 sec excited state (88 kev) is probably $\frac{7}{2}^+ \rightarrow \frac{7}{2}^+$. The fT product for K capture ($T_0/|M_0|^2 \simeq 2{\cdot}5 \times 10^5$ sec) is consistent with an allowed transition.

The recent measurement by Wapstra (1957), $L/K = 0{\cdot}24 \pm 0{\cdot}04$, in agreement with that of der Mateosian ($0{\cdot}23 \pm 0{\cdot}03$) should permit a useful comparison with theory if the energy of the transition were accurately known. But this energy, obtained approximately by the systematics of the energy balance, is $\Delta M_a \simeq 160$ kev. The calculation gives $L/K \simeq 0{\cdot}26$ if one takes $g_{L_I}^2/g_K^2 = 0{\cdot}113$.

Before stating a conclusion it is necessary to know exactly ΔM_a. One could measure, for example, the γ spectrum of internal bremsstrahlung in coincidence with the Ag K capture radiation.

^{125}I

This case is similar to ^{109}Cd. The transition to the 35·4 kev excited state is allowed ($\frac{5}{2}^+ \rightarrow \frac{3}{2}^+$ and $T_0/|M_0|^2 \simeq 1{\cdot}2 \times 10^5$ sec) and the L/K ratio is known accurately: $L/K = (23 \pm 3)\%$ (der Mateosian 1953). Unfortunately $\Delta M_a \simeq 150$ kev is known only approximately, from the systematics of energy balance, and leads to $L/K \simeq 22{\cdot}8\%$ (with $g_{L_I}^2/g_K^2 = 0{\cdot}123$).

It would also be useful to measure ΔM_a accurately using the γ spectrum of internal bremsstrahlung in coincidence with the K radiation following capture.

^{131}Cs

In this case $\Delta M_a = (360 \pm 10)$ kev, obtained from the maximum energy of the γ bremsstrahlung radiation; but the experimental value $L/K \simeq 15\%$ (Fink and Robinson 1955) is not known accurately enough to be compared usefully with theory.

The transition is simple, allowed ($\frac{5}{2}^+ \rightarrow \frac{3}{2}^+$ and $T_0/|M_0|^2 \simeq 3{\cdot}2 \times 10^5$ sec) and the calculation leads to $L/K \simeq 15{\cdot}2\%$ with $g_{L_I}^2/g_K^2 = 0{\cdot}124$.

9.2.2. *Parity-forbidden transitions.*

^{113}Sn

The transition to the excited state corresponds probably to $\frac{1}{2}^+ \rightarrow \frac{1}{2}^-$ ($s_{1/2} \rightarrow p_{1/2}$) and should be parity forbidden. However, the magnitude $T_0/|M_0|^2 = 1{\cdot}28 \times 10^5$ sec is small for such a transition. The value of the transition $\Delta M_a = 490$ kev is well known (from coincidences of γ bremsstrahlung with K radiation of capture), and will enable an interesting comparison with theory when the experimental ratio $L/K \simeq 17\%$ (Avignon 1956) can be measured with more accuracy.

Assuming an allowed transition, the calculation gives a satisfactory approximate value $L/K \simeq 22{\cdot}5\%$. The calculation corresponding to the actual parity-forbidden transition is not useful so long as the experimental value has not been precisely stated.

^{128}I

This is the only case where both $\Delta M_a = 2132$ kev (from β^+ spectrum) and $L/K = 0{\cdot}148$ (Scobie and Gabathuler 1958) are known with accuracy. The transitions to the first and second excited states are probably parity-forbidden,

being $2^- \rightarrow 2^+$. The value of $T_0/|M_0|^2 \simeq 2{\cdot}6 \times 10^7$ sec seems to correspond well with a transition of this type with a high value of the final spin. The experimental result of Scobie and Gabathuler (1958) corresponds to two partial transitions C_1 (27%) and C_2 (4%). The calculation for an allowed transition gives $L/K \simeq 13{\cdot}1\%$ for C_1 and 13·6% for C_2 (taking $g_{L_I}^2/g_K^2 = 0{\cdot}123$, after Brysk). The correction due to the second excited state can be ignored.

The exact theoretical result for a parity-forbidden transition comprises several nuclear matrix elements and interference V–A terms, and is a function of the relative values of these nuclear matrix elements. An approximate value can be obtained if one assumes that the matrix elements corresponding to $T_1(\sigma Y_1)$ and $T_1(Y_1)$ are of the same order of magnitude, while that corresponding to $T_1(\alpha Y_0)$ is smaller by a factor $\nu = \pm \bar{v}/c$ ($\bar{v}$ being the average velocity of the nucleon undergoing the transition in the nucleus); but the sign of ν is not known. The calculation gives finally L/K varying from 12% to 15% according to the sign of ν; let us note that this L/K ratio calculated for the actual forbidden transition is very close to that of the allowed transition.

The value of $T_0/|M_1|^2$ obtained for the C_1 forbidden transition, between 7×10^6 and 2×10^7 sec according to the sign of ν, seems large and gives an important nuclear reduction factor ($R \simeq 10^{-4}$).

^{235}Np

The energy available in this transition is not known, so the calculation of L/K cannot be a test for the theory. On the contrary, the experimental measurement of Gindler *et al.* (1958), $L/K \simeq 36{\cdot}7$, makes possible the calculation of $\Delta M_a = 126$ kev.

This value does not differ much from those of Hoff (1956), $\Delta M_a = 130$ kev and Gindler *et al.* (1958), $\Delta M_a = 123$ kev, obtained assuming an allowed transition. But it is clearly smaller than the result of Glass *et al.* (1955), $\Delta M_a = 170$ kev, obtained from the disintegration cycle

$$^{235}\text{Np} \xrightarrow{\alpha} \text{Pa} \xrightarrow{\beta} \text{Th} \xrightarrow{\alpha} \text{U} \xrightarrow{K} {}^{235}\text{Np}.$$

Finally, let us note recent measurements by Perrin (1958), by means of X–γ coincidences, of L/K ratios for capture going to excited states of ^{65}Zn, ^{75}Se, ^{101}Rh, ^{114}In, ^{152}Eu. This method is very interesting, but the results are still not accurate enough to enable a comparison with theory to be made.

9.2.3. *Conclusion.*

All the above results show that there is no important disagreement between the experimental and the theoretical values. The deviations observed earlier by Langevin (1955 a) and Radvanyi (1954) on ^{71}Ge and ^{79}Kr are largely reduced by progress in methods of measurement, and by a better knowledge of the fluorescence yields. The present theory of capture, as well as the values of the radial functions of Brysk and Rose (1955) seem satisfactory. But there are still only very few transitions for which accurate values of L/K (independent of R_K) and of ΔM_a are available.

Let us note also the important difference in the ^{37}A capture between the number of K Auger electrons (65%) measured by Kofoed-Hansen (1954) and the number which can be estimated from the Auger effect $(K(1-R_K)/(L+K) \simeq 84\%)$. If this

experimental result is confirmed, it could be necessary to modify the present theory of the Auger effect to make allowance, as Robinson and Fink (1955) comment, for the position correlation between the electrons in the atom (Daudel 1955).

§ 10. Capture Positron Ratio

β^+ emission is always accompanied by electron capture; the measurement of C/β^+ branching ratio is necessary to know the decay scheme of the β^+ emitter.

The comparison of the measured value with the theoretical estimate has a great interest chiefly for allowed and 'unique' forbidden transitions $\Delta I = n+1$, $\Delta\pi = (-1)^n$. In these types of transitions the nuclear matrix elements in the C/β^+ ratio disappear if one does not take into account the Fierz (1937) interference terms. This comparison is then a good test of the validity of β disintegration theory and of the eigenfunctions of the electron.

The measurements of C/β^+ ratios are now accurate enough to determine the wave functions of the captured electron and of the emitted positron, taking into account the secondary effects such as the finite size of the nucleus and the screening effect of the other atomic electrons; it is even sometimes necessary to take into account the secondary terms of the series expansion of positron functions. This correction is useful for large values of transition energy corresponding to a small value of the branching ratio. It is of course necessary to take L capture into account in the comparison with theory, which is made on the basis of the K/β^+ ratio.

The β interaction being supposed (V, A) there are no Fierz terms from the (A, T) or (S, V) interactions; and K/β^+ is independent of the nuclear matrix element for the allowed transitions and for the 'unique' transitions.

10.1. K/β^+ *Ratio for Allowed Transitions*

10.1.1. *Theoretical results.*

The transition probability for capture is given by Møller's formula. For β^+ emission the probability is given by Fermi's formula

$$P_+(Z,\epsilon)\,d\epsilon = \frac{g^2}{2\pi^3}|M|^2\frac{1+\gamma_0}{2}\epsilon p q^2 F_0^+(Z,\epsilon)\,d\epsilon$$

in which the nuclear matrix element M is the same as that (56) given for capture, $F_0^+(Z,\epsilon)$ is Fermi's function for allowed transitions; the correction due to the Coulomb field acting on the positrons is normalized so that $F_0^+(Z=0)=1$. This F function is given in zero order approximation by the table of the National Bureau of Standards (Feister 1952) and has been calculated by Dzelepov and Zirianova (1952) with second order corrections. The $(1+\gamma_0)/2$ factor is the term L_0 of Greuling (1942). Taking into account the formulae of Møller and Fermi, the K/β^+ ratio is written

$$\frac{K}{\beta^+} = \frac{\pi}{2}\frac{(\epsilon_0+\epsilon_K)^2 g_K^2}{\frac{1}{2}(1+\gamma_0)\int_1^{\epsilon_0}\epsilon p q^2 F_0^+(Z,\epsilon)\,d\epsilon}.$$

The radial functions for the bound electron have been calculated by Brysk and Rose (1955, 1958) taking into account screening, nuclear radius corrections, and the variation of the wave function over the nuclear volume. For the free electron

the F functions of Feister or Dzelepov must be corrected for screening effect calculated by Reitz (1950) and for nuclear radius calculated by Rose and Holmes (1951) : this nuclear radius effect is about 1% and can be neglected.

Further, it is assumed that the β disintegration takes place close to the surface of the nucleus ($r < R$). One can say in support of this hypothesis that the transitions of deep nucleons are forbidden by Pauli's principle : a transition $n \rightleftharpoons p$ of a nucleon from a deep filled shell $r < R$ can take place only with an appreciable variation of the angular momentum and of the energy. Such a transition cannot exist in β disintegration, since the available energy is not large enough and an appreciable variation of the angular momentum produces an appreciable decrease of the probability (pR and $qR \ll 1$). Let us note that it is not the same for μ capture, since here the available energy is appreciable (about 100 MeV) and because an appreciable variation of the angular momentum does not decrease the probability ($qR \simeq 1$). These remarks enable us to consider the transition $n \rightleftharpoons p$ as taking place preferably near the surface. Brysk and Rose (1955, 1958) have given the radial functions of the bound electron, assuming that the capture is spread uniformly in the nucleus. They obtain then for g_K^2 a difference of no more than about 2%. The truth is certainly between these two approximations, whence we may possibly assume a correction of no more than about 1%.

10.1.2. *Experimental results—comparison with theory.*

The direct measurement of the relative probability of capture consists in a measurement of the intensity of the photons or the Auger electrons accompanying capture. The intensity measurement of positrons is made from the β^+ spectrum or the annihilation γ radiation. When capture or β^+ emission is followed by γ transition, the measurement of the total intensity of γ photons (or eventually of internal conversion electrons) gives the total probability ($C + \beta^+$) ; by measuring in addition the coincidences between the photons γ (or the internal conversion electrons) and the radiations of the transition $C + \beta^+$ (K photons, positrons or annihilation photons) one can obtain the K/β^+ ratio. Present-day development of scintillation spectrometry, of proportional counters and methods of coincidence have considerably improved the sensitivity and the accuracy of the measurements. However the values observed for K/β^+ (table 6) still show a considerable scatter.

The basic work is covered by Feenberg and Trigg's (1950) curves, Zweifel's (1957) tables and the values calculated by the group of Brookhaven (Perlman *et al.* 1958).

In these values all the corrections which are now known were not taken into account, and we have therefore calculated the results again using Brysk and Rose's curves (1955) for the radial functions of the bound electron and Dzelepov's tables for β^+ emission. We have also taken into account for β^+ emission the screening effect.

10.2. K/β^+ *Ratio for Forbidden Transitions*

10.2.1. *Forbidden 'unique' transitions* $\Delta I = n+1, \Delta\pi = (-1)^n$.

The β^+ emission probability is given by Nataf and Bouchez (1952)

$$P_+(Z, \epsilon)\, d\epsilon = \frac{g^2}{2\pi^3} |M_{n+1}|^2 \frac{(2n+1)!}{[(2n+1)!!]^2} \epsilon p q^2 \sum_{r+s=n} \frac{(pR)^{2r}}{(2r+1)!} \cdot \frac{(qR)^{2s}}{(2s+1)!} \cdot \frac{\gamma_r + r + 1}{2r+2} F_r$$

Table 6. K/β^+ Ratio for Allowed Transitions—Experimental and Theoretical Values

Transition	W_0(kev)	K/β^+ Observed		K/β^+ Calculated	
$^{18}F(1^+) \rightarrow 0^+$	649	0.030 ± 0.002	Drever *et al.* 1956	0·029	Perlman *et al.* 1958
				0·029	Nguyen-Khac 1959
		0.103 ± 0.006	Sherr and Miller 1954		
		0.065 ± 0.009	Charpak 1955		
$^{22}Na(3^+) \rightarrow 2^+$	541	0.122 ± 0.010	Allen *et al.* 1955	0·104	Zweifel 1954
		0.102 ± 0.008	Konijn *et al.* 1958	0·107	Nguyen-Khac 1959
		0.067 ± 0.016	Blue and Bleuler 1955	0·045	Perlman *et al.* 1958
$^{44}Sc(2^+) \rightarrow 2^+$	1467	0.021 ± 0.019	Konijn *et al.* 1958	0·042	Nguyen-Khac 1959
		0.70 ± 0.09	Bock 1955	0·73	Perlman *et al.* 1958
$^{46}V(4^+) \rightarrow 4^+$	695	0.68 ± 0.02†	Konijn *et al.* 1958	0·66	Nguyen-Khac 1959
		0.69 ± 0.03‡			
		1.84 ± 0.22	Sehr 1954	1·9	Perlman *et al.* 1958
$^{52}Mn(6^+) \rightarrow 6^+$	575	1.81 ± 0.07†	Konijn *et al.* 1958	1·77	Nguyen-Khac 1959
		1.86 ± 0.09‡			
		5.4 ± 0.2	Cook and Tomnovec 1956		
$^{54}Co(2^+) \rightarrow 2^+$	472	6.1 ± 0.8	Grace *et al.* 1956	5·2	Perlman *et al.* 1958
		5.2 ± 0.16	Konijn *et al.* 1958	4·9	Nguyen-Khac 1959
		0.18 ± 0.03	Huber *et al.* 1949	0·29	Perlman *et al.* 1958
$^{61}Cu(3/2^-) \rightarrow 3/2^-$	1205	0.22 ± 0.03	Bouchez 1952	0·28	Nguyen-Khac 1959
		1.75 ± 0.2	Huber *et al.* 1949		
		2.6 ± 0.3	Bouchez and Kayas 1949	2·3	Perlman *et al.* 1958
$^{64}Cu(1^+) \rightarrow 0^+$	657	2.32 ± 0.28	Reynolds 1950	2·07	Nguyen-Khac 1959
		$1.8 \quad -2.07$	Plassman and Scott 1951		
		30	Perrin 1958		
		28.0 ± 3.2	Perkins and Haynes, 1953	30	Zweifel 1958
$^{65}Zn(5/2^-) \rightarrow 3/2^-$	325	26 ± 3	Avignon 1956	30·7	Nguyen-Khac 1959
		25 ± 2	Gleason 1959		
		2·7	Goldhaber *et al.* 1951		
$^{95}Zr(9/2^+) \rightarrow 9/2^+$	901	3·7	Shore *et al.* 1953	2·8	Perlman *et al.* 1958
				310	Perlman *et al.* 1958
$^{107}Cd(5/2^+) \rightarrow 7/2^+$	320	320 ± 30	Bradt *et al.* 1945	331	Nguyen-Khac 1959
$^{111}Sn(7/2^+) \rightarrow 9/2^+$	1510	2.5 ± 0.25	McGinnis 1951	1·5	Perlman *et al.* 1958

† proportional counter. ‡ β spectrometer.

with $\gamma_r^2 = (r+1)^2 - \alpha^2 Z^2$, and F_r the Coulomb correction factor ($F_r = 1$ for $Z = 0$) which generalize the Fermi function F_0 for the allowed transitions. The nuclear matrix element is the same as that of capture probability given by

$$P_K = \frac{g^2}{4\pi^2} |M_{n+1}|^2 \frac{1}{[(2n+1)!!]^2} q^2 (qR)^{2n} g_K^2.$$

Transition type: ^{122}Sb ($\Delta J = 2$, yes).

In this case $n = 1$ the K/β^+ ratio is written

$$\frac{K}{\beta^+} = \frac{\pi}{2} \frac{q_K^4 g_K^2}{\int_1^{\epsilon_0} [\frac{1}{2}(1+\gamma_0) F_0 q^2 + \frac{1}{4}(2+\gamma_1) F_1 p^2)] \epsilon p q^2 d\epsilon}$$

or using the Greuling (1942) notation $\begin{cases} L_0 \equiv \dfrac{1+\gamma_0}{2} \\ L_1 \equiv \dfrac{F_1}{F_0} \dfrac{p^2}{9} \dfrac{2+\gamma_1}{4} \end{cases}$

$$\frac{K}{\beta^+} = \frac{\pi}{2} \frac{q^4 g_K^2}{\int_1^{\epsilon_0} \epsilon p q^2 L_0 F_0 \left(q^2 + 9 \dfrac{L_1}{L_0}\right) d\epsilon}.$$

10.2.2. K/β^+ *ratio for first-forbidden transitions.*

For the other forbidden transitions, even if one does not take Fierz's terms into account, there exist different nuclear matrix elements and the value of the K/β^+ ratio depends on their relative value.

For the first forbidden transitions one must distinguish the $\Delta J = 0$ and the $\Delta J = 1$ transitions.

10.3. *Conclusion*

The theoretical values for allowed transitions calculated by Nguyen-Khac, in this laboratory, with Dzelepov's tables and Brysk and Rose's functions do not differ greatly from the values of other authors and agree in general with experiment. The agreement is within 5% in most cases. Perhaps the case of ^{44}Sc has to be reconsidered. For the forbidden ' unique ' transitions, the measurement on ^{84}Rb (Welker and Perlman 1955) is in disagreement with the calculated value by Perlman (1958).

Of course it would be desirable to improve both experiment and calculation but it can be said that the situation for K/β^+ ratios is fairly good.

In the case of ' non-unique ' transitions where many matrix elements are involved, it seems that the actual K/β^+ ratio is significantly higher than the ' allowed ratio '. This could be explained by assuming that the tensor T_2 (σY_1) gives an appreciable contribution (Perlman, Welker and Wolfsberg 1958).

Table 7. K/β^+ Ratio for Forbidden 'Unique' Transitions ($\Delta J = 2$, yes)

Transition	W_0(kev)	K/β^+ Observed		K/β^+ Calculated $\Delta J = 2$, yes	Allowed	References
^{90}Rb $(2^-) \rightarrow 0^+$	1700	$2{\cdot}06 \pm 0{\cdot}36$	Welker 1955	0·83	0·26	Perlman *et al.* 1958
^{122}Sb $(2^-) \rightarrow 0^+$	565	300 ± 50	Perlman *et al.* 1958	275 ± 60	49	Perlman *et al.* 1958
				285	48·6	Nguyen-Khac 1959
^{126}I $(2^-) \rightarrow 0^+$	1155	$18 \pm 0{\cdot}3$	Perlman *et al.* 1958	$17{\cdot}3 \pm 1$	4·6	Perlman *et al.* 1958

Table 8. K/β^+ Ratio for First Forbidden Transitions with $\Delta J = 0$. (*After* Perlman *et al.* 1958)

Transition	W_0(kev)	K/β^+ Observed	K/β^+ Calculated $\Delta J = 0$, yes	Allowed
^{74}As $(2^-) \rightarrow 2^+$	920	1·5	1·47	1·17
^{84}Rb $(2^-) \rightarrow 2^+$	782	$5{\cdot}1 \pm 0{\cdot}4$	$4{\cdot}2 \pm 0{\cdot}4$	$3{\cdot}4 \pm 0{\cdot}3$
^{126}I $(2^-) \rightarrow 2^+$	470	148	143	122

References

Abelson, P., 1939, *Phys. Rev.*, **54**, 424.

Allen, R. A., Burcham, W. E., Chackett, K. F., Munday, G. L., and Reasbeck, P., 1955, *Proc. Phys. Soc.* A, **68**, 681.

Alvarez, L. W., 1937, *Phys. Rev.*, **52**, 134 (L); 1938, *Ibid.*, **54**, 486.

Alvarez, L. W., Helmholtz, A. C., and Wright, B. T., 1941, *Phys. Rev.*, **60**, 160.

Auger, P., 1925, *J. Phys. Radium*, **6**, 205; 1926, *Ann. Phys., Paris*, **6**, 183.

Avignon, P., 1953, *J. Phys. Radium*, **14**, 637; 1956, *Ann. Phys., Paris*, **1**, 10.

Avignon, P., Michalowicz, A., and Bouchez, R., 1955, *J. Phys. Radium*, **16**, 404.

Awscholom, M., 1956, *Phys. Rev.*, **101**, 1041.

Bellicard, J. B., and Moussa, A., 1957, *J. Phys. Radium*, **18**, 115.

Benoist, P., 1949, *C.R. Acad. Sci., Paris*, **228**, 309.

Benoist, P., and Jean, M., 1949, *C.R. Acad. Sci., Paris*, **228**, 1848.

Benoist-Gueutal, P., 1953, *Thèse*, University of Paris, No. 3388.

Bergström, I., 1955, *β and γ Ray Spectroscopy*, p. 624. Ed. K. Siegbahn (Amsterdam: North-Holland).

Bergström, I., and Hill, R. D., 1954, *Ark. Fys.*, **8**, 21.

Bethe, H. A., 1933, *Handb. d. Phys.*, **24/1**, 301 (Berlin: Springer).

Bethe, H. A., and Bacher, R. F., 1936, *Rev. Mod. Phys.*, **8**, 82.

Bisi, A., Germagnoli, E., and Zappa, L., 1955, *Nuovo Cim.*, **11**, 1052.

Bisi, A., Germagnoli, E., Zappa, L., and Zimmer, E., 1955, *Nuovo Cim.*, **2**, 290.

Blatt, J. M., and Weisskopf, V. F., 1952, *Theoretical Nuclear Physics* (New York: Wiley).

Blue, J. W., and Bleuler, E., 1955, *Phys. Rev.*, **100**, 1324.

Bock, R., 1955, *Nuclear Level Schemes* (Washington: United States Atomic Energy Commission).

Boehm, F., and Wapstra, A. H., 1957, *Phys. Rev.*, **107**, 1462.

Bouchez, R., 1950, *Thèse*, University of Paris, No. 2340; 1952, *Physica*, **18**, 1171.

Bouchez, R., Daudel, P., Daudel, R., Muxart, R., and Rogozinski, A., 1949, *J. Phys. Radium*, **10**, 201.

Bouchez, R., Hubert, P., Perrin, N., and Sakai, M., 1953, *J. Phys. Radium*, **14**, 295.

Bouchez, R., and Kayas, G., 1949, *J. Phys. Radium*, **10**, 110.

Bouchez, R., Michalowicz, A., and Avignon, P., 1955, *J. Phys. Radium*, **16**, 404.

Bouchez, R., and Nataf, R., 1952, *C.R. Acad. Sci., Paris*, **234**, 86; 1953, *J. Phys. Radium*, **14**, 217.

Bouchez, R., Robert, J., and Tobailem, J., 1953, *J. Phys. Radium*, **14**, 281.

Bouchez, R., Tobailem, J., Robert, J., Muxart, R., Meller, R., and Daudel, P. R., 1956, *J. Phys. Radium*, **17**, 363.

Bradt, H., Gugelot, P. C., Huber, O., Medicus, H., Preiswerk, P., and Scherrer, P., 1945 a, *Helv. Phys. Acta*, **18**, 351; 1945 b, *Phys. Rev.*, **68**, 57.

Brysk, H., and Rose, M. E., 1955, *O.R.N.L. Report*, No. 1830; 1958, *Rev. Mod. Phys.*, **30**, 1169.

Burford, A. O., Perkins, J. F., and Haynes, S. K., 1955, *Phys. Rev.*, **99**, 3.

Burgy, M. T., Krohn, V. E., Novey, T. B., Ringo, G. R., and Telegdi, V. A., 1959, *Congrès International de Physique Nucléaire* 1958 (Paris: Dunod).

Burhop, E. H. S., 1952, *The Auger Effect* (Cambridge: University Press); 1955, *J. Phys. Radium*, **16**, 625.

Cauchois, Y., and Hulubei, H., 1947, *Longueur d'onde des émissions* X *et des discontinuités d'absorption* (Paris: Hermann).

Charpak, G., 1955, *J. Phys. Radium*, **16**, 62.

Cohen, S. G., 1951, *Nature, Lond.*, **167**, 779.

Compton, A. H., and Allison, S. K., 1935, *X-rays in Theory and Experiment* (New York: Van Nostrand).

Cook, C. S., 1951, *Amer. J. Phys.*, **19**, 37.

Cook, C. S., and Langer, L. M., 1948, *Phys. Rev.*, **74**, 1241 (A).

Cook, C. S., and Tomnovec, F. M., 1956, *Phys. Rev.*, **104**, 1407.

Coster, D., and Kronig, L., 1935, *Physica*, **2**, 13.

COWAN, C. L., HARRISON, F. B., LANGER, L. M., and REINES, F., 1956, *Nuovo Cim.*, **3**, 649.
COWAN, C. L., REINES, F., and HARRISON, F. B., 1954, *Phys. Rev.*, **96**, 1294.
CULLIGAN, G., FRANK, S. G. F., HOLT, J. R., KLUYVER, J. C., and MASSAM, T., 1957, *Nature, Lond.*, **180**, 751.
CURIE, I., and JOLIOT, F., 1933, *J. Phys. Radium*, **4**, 278.
DAUDEL, R., 1955, *J. Phys. Radium*, **16**, 503.
DAUDEL, R., BENOIST, P., JACQUES, R., and JEAN, M., 1947, *C.R. Acad. Sci., Paris*, **224**, 1427.
DAVIS, R., JR., 1956, *Bull. Amer. Phys. Soc.*, Ser. 2, **1**, 219.
DREVER, R. W. P., and MOLJK, A., 1957, *Phil. Mag.*, **2**, 427.
DREVER, R. W. P., MOLJK, A., and SCOBIE, J., 1956, *Phil. Mag.*, **1**, 942.
DZELEPOV, B. S., and ZIRIANOVA, L. N., 1952, *Tables pour la désintégration β*. Ed. de l'Université de Léningrad.
EMMERICH, W. S., SINGER, S. E., and KURBATOV, J. D., 1954, *Phys. Rev.*, **94**, 113.
FAIRBROTHER, J. A. V., PARKYN, D. G., and O'CONNOR, B. M., 1957, *Proc. Phys. Soc.* A, **70**, 262.
FARRELLY, B., KOERTS, L., BENCZER, N., VAN LIESHOUT, R., and WU, C. S., 1955, *Phys. Rev.*, **99**, 1440.
FEATHER, N., 1952 a, *Proc. Roy. Soc. Edinb.* A, **63**, 242 ; 1952 b, *Physica*, **13**, 1023 ; 1952 c, *Nuclear Stability Rules* (Cambridge : University Press).
FEENBERG, E., and TRIGG, G., 1950, *Rev. Mod. Phys.*, **22**, 399.
FEISTER, I., 1952, *Tables for the Analysis of β Spectra* (Washington : N.B.S. Applied Mathematics series No. 13).
FERMI, E., 1934, *Z. Phys.*, **88**, 161.
FEYNMAN, R. P., and GELL-MANN, M., 1958, *Phys. Rev.*, **109**, 193.
FIERZ, M., 1937, *Z. Phys.*, **104**, 533.
FINK, R. W., and ROBINSON, B. L., 1955, *Phys. Rev.*, **93**, 1293.
FLACK, F. C., and MASON, P., 1958, *Proc. Phys. Soc.*, **71**, 247.
FRAUENFELDER, H., BOBONE, R., VON GOELER, E., LEVINE, N., LEWIS, H. R., PEACOCK, R. N., ROSSI, A., and DE PASQUALI, G., 1957, *Phys. Rev.*, **106**, 386 (L).
FRIEDMAN, J. I., and TELEGDI, V. L., 1957, *Phys. Rev.*, **105**, 1681 (L).
GARWIN, R. L., LEDERMAN, L. M., and WEINRICH, M., 1957, *Phys. Rev.*, **105**, 1415 (L).
GINDLER, J. E., HUIZENGA, J. R., and ENGELKEMEIR, D. W., 1958, *Phys. Rev.*, **109**, 1263.
GLASS, R. A., THOMPSON, and SEABORG, G. T., 1955, *J. Inorg. Nuclear Chem.*, **1**, 3.
GLAUBMANN, M. J., 1955, *Phys. Rev.*, **98**, 645.
GLEASON, G. I., 1959, *Phys. Rev.*, **113**, 287.
GOLDHABER, M., DER MATEOSIAN, E., SCHARFF-GOLDHABER, G., and SUNYAR, A. W., 1951, *Phys. Rev.*, **83**, 661.
GOLDHABER, M., GRODZINS, L., and SUNYAR, A. W., 1958, *Phys. Rev.*, **109**, 1015 (L).
GRACE, M. A., JONES, G. A., and NEWTON, J. O., 1956, *Phil. Mag.*, **1**, 363.
GREULING, E., 1942, *Phys. Rev.*, **61**, 568.
HAMILTON, D. R., ALFORD, W. P., and GROSS, L., 1953, *Phys. Rev.*, **92**, 1521.
HAYNES, S. K., and ACHOR, W. T., 1955, *J. Phys. Radium*, **16**, 635.
HAYWARD, R. W., and HOPPES, D. D., 1955, *Phys. Rev.*, **99**, 659 (A).
HEINTZE, J., 1954, *Z. Naturf.*, **9a**, 460.
HILL, R. D., CHURCH, E. L., and MIHELICH, J. W., 1952, *Rev. Sci. Instrum.*, **23**, 523.
HOFF, R. W., OLSEN, J. L., and MANN, L. G., 1956, *Phys. Rev.*, **102**, 805.
HUBER, O., RUETSCHI, R., and SCHERRER, P., 1949, *Helv. Phys. Acta*, **22**, 375.
JACOBSEN, J. C., 1937, *Nature, Lond.*, **139**, 879.
KOFOED-HANSEN, O., 1954, *Phys. Rev.*, **96**, 1045.
KONIJN, J., VAN NOOIJEN, B., HAGEDOORN, H. L., and WAPSTRA, A. H., 1958, *Nucl. Phys.*, **9**, 296.
KONOPINSKI, E. J., 1943, *Rev. Mod. Phys.*, **15**, 209 ; 1958, *Proc. Rehovoth Conference*, 319 (Amsterdam : North-Holland).
KONOPINSKI, E. J., and UHLENBECK, G. E. 1941, *Phys. Rev.*, **60**, 308.
KRAUSHAAR, J. J., WILSON, E. D., and BAINBRIDGE, K. T., 1953, *Phys. Rev.*, **90**, 610.
LABERRIGUE-FROLOW, J., and RADVANYI, P., 1956, *J. Phys. Radium*, **17**, 944.

LANDAU, L., 1957, *Nucl. Phys.*, **3**, 127.
LANGER, L. M., and MOFFAT, R. J. D., 1952, *Phys. Rev.*, **88**, 689.
LANGEVIN, M., 1954, *C.R. Acad. Sci., Paris*, **238**, 1518 ; 1955 a, *Thèse*, University of Paris, No. 3656 ; 1955 b, *J. Phys. Radium*, **16**, 516 ; 1955 c, *C.R. Acad. Sci., Paris*, **241**, 33.
LEE, T. D., and YANG, C. N., 1956, *Phys. Rev.*, **104**, 254 ; 1957, *Ibid.*, **105**, 1671.
LEININGER, R. F., SEGRÈ, E., and WIEGAND, C., 1949, *Phys. Rev.*, **76**, 897.
LINDQVIST, T., and WU, C. S., 1955, *Phys. Rev.*, **100**, 145.
LIUBIMOV, V. A., 1959, *Congrès International de Physique Nucléaire* 1958 (Paris : Dunod).
MCCUE, J. J. G., and PRESTON, W. M., 1951, *Phys. Rev.*, **84**, 384.
MCGINNIS, C. L., 1951, *Phys. Rev.*, **81**, 734.
MAJOR, J. K., and BIEDENHARN, L. C., 1954, *Rev. Mod. Phys.*, **26**, 321.
MAJORANA, E., 1937, *Nuovo Cim.*, **14**, 171.
MALCOLM, I., 1952, *Phil. Mag.*, **43**, 1011.
MALCOLM, I., and STRACHAN, C., 1951, *Proc. Camb. Phil. Soc.*, **47**, 610.
MARION, J. B., and KAVANAGH, R. W., 1956, *Phys. Rev.*, **104**, 107.
MARSHAK, R. E., 1942, *Phys. Rev.*, **61**, 431.
DER MATEOSIAN, E., 1953, *Phys. Rev.*, **92**, 938.
MICHALOWICZ, A., 1956, *Thèse*, University of Paris, No. 3780.
MILLER, M. M., and WILKINSON, R. G., 1951, *Phys. Rev.*, **83**, 1050.
MØLLER, C., 1937 a, *Phys. Z. Sowjet.*, **11**, 9 ; 1937 b, *Phys. Rev.*, **51**, 84.
MUKLIN, A. I., OZEROV, E. B., PONTECORVO, B., 1958, *J.I.N.R.* (*Doubna*) *P*, **159**.
NATAF, R., 1954, *C.R. Acad. Sci., Paris*, **238**, 1012, 1117, **239**, 341 ; 1956, *J. Phys. Radium*, **17**, 480 ; 1957, *C.R. Acad. Sci., Paris*, **244**, 884, 1031, 1187, 1917 ; 1959, *Congrès International de Physique Nucléaire* 1958 (Paris : Dunod).
NATAF, R., and BOUCHEZ, R., 1952, *J. Phys. Radium*, **13**, 81, 190.
NGUYEN-KHAC, U., 1960, *Thèse*, University of Grenoble, No. 136.
PAULI, W., 1933 a, *7ème Conseil Solvay, Bruxelles* ; 1933 b, *Handb. d. Phys.*, **24**, 1225 (Berlin : Springer) ; 1955, *Niels Bohr and the Development of Physics* (London : Pergamon Press).
PERKINS, J. F., and HAYNES, S. K., 1953, *Phys. Rev.*, **92**, 687.
PERLMAN, M. L., WELKER, J. P., and WOLFSBERG, M., 1958, *Phys. Rev.*, **110**, 381.
PERRIN, N., 1953, *J. Phys. Radium*, **14**, 273 ; 1958, *Thèse*, University of Paris, No. 4058.
PINCHERLE, L., 1935, *Nuovo Cim.*, **12**, 81.
PLASSMAN, E., and SCOTT, F. R., 1951, *Phys. Rev.*, **84**, 156.
PONTECORVO, B., 1946, *Rep. Progr. Phys.*, **11**, 32 (London : Physical Society).
PONTECORVO, B., KIRKWOOD, D. H. W., and HANNA, G. C., 1949, *Phys. Rev.*, **75**, 982.
PURSEY, D. L., 1951, *Phil. Mag.*, **42**, 1193.
DE RAAD, B., MIDDELKOOP, W. C., VAN NOOYEN, B., and ENDT, P. M., 1954, *Physica*, **20**, 1278.
RADVANYI, P., 1954, *Thèse*, University of Paris, No. 3572 ; 1955, *J. Phys. Radium*, **16**, 509.
REITZ, J., 1950, *Phys. Rev.*, **77**, 10.
REYNOLDS, J. H., 1950, *Phys. Rev.*, **79**, 789.
RICHARDS, H. T., SMITH, R. V., and BROWNE, C. P., 1950, *Phys. Rev.*, **80**, 524.
RIDLEY, B. W., 1956, *Progress in Nuclear Physics*, **5**, 188.
RIETJENS, L. H. T., 1956, *Thesis*, University of Utrecht.
RIETJENS, L. H. T., VAN DEN BOLDT, H. J., and ENDT, P. M., 1954, *Physica*, **20**, 107.
ROBINSON, B. L., and FINK, R. W., 1955, *Rev. Mod. Phys.*, **27**, 424.
ROSE, M. E., 1937, *Phys. Rev.*, **51**, 484 ; 1951, *Ibid.*, **82**, 389 ; 1955, *Multipole Fields* (London : Chapman and Hall).
ROSE, M. E., and HOLMES, D. K., 1951, *Phys. Rev.*, **83**, 190.
ROSE, M. E., and OSBORN, R. K., 1954, *Phys. Rev.*, **93**, 1315, 1326.
ROSE, M. E., and PERRY, C. L., 1953, *Phys. Rev.*, **90**, 479.
ROSENBLUM, S., and GUILLOT, M., 1937, *C.R. Acad. Sci., Paris*, **204**, 1727.
ROSENBLUM, S., SANT'ANA DIONISIO, J., and VALADARES, M., 1956, *J. Phys. Radium*, **17**, 112.
RUBENSTEIN, R. A., and SNYDER, J. N., 1955, *Phys. Rev.*, **97**, 1653.
SALAM, A., 1957, *Nuovo Cim.*, **5**, 299.
SARAF, B., 1954, *Phys. Rev.*, **94**, 642.

SARGENT, B. W., 1933, *Proc. Roy. Soc.* A, **139**, 659.
SCHOENFELD, W. A., DUBORG, R. W., PRESTON, W. M., and GOODMAN, C., 1952, *Phys. Rev.*, **85**, 873.
SCHOPPER, H., 1957, *Phil. Mag.*, **2**, 710.
SCOBIE, J., and GABATHULER, E., 1958, *Proc. Phys. Soc.*, **72**, 437.
SEHR, R., 1954, *Z. Phys.*, **137**, 523.
SHERR, R., and MILLER, R. H., 1954, *Phys. Rev.*, **93**, 1076.
SHORE, F. J., BLENDEL, W. L., BROWN, H. N., and BECKER, R. A., 1953, *Phys. Rev.*, **91**, 1203.
SIZOO, G. J., 1937, *Physica*, **4**, 468.
SLATER, J. C., 1930, *Phys. Rev.*, **36**, 57.
SNELL, A. H., and PLEASONTON, F., 1955, *Phys. Rev.*, **100**, 1396.
STELSON, P. H., PRESTON, W. M., and GOODMANN, C., 1950, *Phys. Rev.*, **80**, 287 (L).
SUZOR, F., and CHARPAK, G., 1959, *J. Phys. Radium*, **20**, 462.
THOMPSON, S. G., 1949, *Phys. Rev.*, **76**, 319 (L).
TRAIL, C. R., and JOHNSON, C. H., 1953, *Phys. Rev.*, **91**, 474.
WAPSTRA, A. H., 1955, *Physica*, **21**, 367 ; 1957, *Nucl. Phys.*, **4**, 325.
WAY, K., KING, R. W., MCGINNIS, C. L., and VAN LIESHOUT, R., 1955, *Nuclear Level Schemes Nat. Res.* TID 5300.
WELKER, J. P., and PERLMAN, M. L., 1955, *Phys. Rev.*, **100**, 74.
WU, C. S., 1959, *Congrès International de Physique Nucléaire* 1958 (Paris : Dunod).
WU, C. S., AMBLER, E., and HAYWARD, R. W., 1957, *Phys. Rev.*, **105**, 1413 (L).
YUKAWA, H., and SAKATA, S., 1935, *Proc. Phys. Math. Soc. Japan*, **17**, 467 ; 1936, *Ibid.*, **18**, 128 ; 1937, *Phys. Rev.*, **51**, 677 (L).
ZWEIFEL, P. F., 1954, *Phys. Rev.*, **96**, 1572 ; 1957, *Ibid.*, **107**, 329 ; 1958, *Proceedings of the Rehovoth Conference on Nuclear Structure* (Amsterdam : North-Holland).

PRECISION MEASUREMENT IN GAMMA-RAY SPECTROSCOPY

By G. A. BARTHOLOMEW, J. W. KNOWLES AND G. E. LEE-WHITING
Atomic Energy of Canada Limited, Chalk River, Ontario, Canada

CONTENTS

	PAGE
§ 1. Introduction	454
§ 2. Spectrometers using coherent scattering	459
2.1. Introduction	459
2.2. Line source spectrometers	466
2.2.1. The line source transmission spectrometer	466
2.2.2. The line focus reflection spectrometer	472
2.3. Broad source transmission spectrometers	473
2.3.1. The flat crystal diffraction spectrometer	473
2.3.2. The Cauchois spectrometer	479
2.4. A discussion of crystal spectrometers	481
2.4.1. Comparison of crystal spectrometers	481
2.4.2. The improvement of spectrometer efficiency	483
2.4.3. The precision of energy measurements	486
§ 3. Spectrometers using the photoelectric effect	487
3.1. Introduction	487
3.2. Lens spectrometers	490
3.2.1. Principles of focusing	490
3.2.2. The Pasadena ring-focus spectrometer	493
3.3. Flat spectrometers	495
3.3.1. Principles of focusing	495
3.3.2. The Chalk River $\pi\sqrt{2}$ spectrometer	499
3.4. Conclusion	502
§ 4. Spectrometers using the Compton effect	504
4.1. Introduction	504
4.2. Development of the flat Compton spectrometer	507
4.3. General properties of Compton spectrometers	510
4.4. Proposals for improved Compton spectrometers	513
4.4.1. Two-directional focusing	513
4.4.2. Use of the back-scattered photon	516
§ 5. Spectrometers using the pair effect	516
5.1. Introduction	516
5.2. Lens-type pair spectrometers	518
5.2.1. Off-axis separation	518
5.2.2. Statistical separation	519
5.3. Flat field 180° pair spectrometer	522
5.3.1. Theory	522
5.3.2. Practical geometries	530
5.3.3. Precision	533
5.4. Comparison of pair spectrometers	533
5.4.1 Acceptance solid angle and transmission	534
5.4.2. The efficiency function	534
5.4.3. Background	535
§ 6. Summary	535
Acknowledgments	541
References	541

Abstract. Gamma-ray spectrometers suitable for precision measurements, i.e. those capable of an energy resolution exceeding 1 per cent and an energy accuracy better than 0·2 per cent, are discussed according to the γ-ray interaction used: coherent scattering (crystal diffraction), photoelectric effect, Compton effect, and pair production. Leading instruments of each type are described and the energy range, resolution, efficiency function, and precision in energy and intensity measurement are discussed and, in some cases, ways of improving existing instruments are indicated. The suitability of the various instruments for different applications is considered and, where more than one type of instrument can be used under identical conditions, attempts are made to compare their relative merits.

§ 1. Introduction

Gamma-ray spectroscopy concerns the determination of the energies and intensities of individual γ-rays often in the presence of a complex spectrum of other γ-rays or a background of extraneous radiation. Precise measurement is facilitated if the spectrometer to be used possesses good energy resolution (full width at half maximum of the γ-ray line). High resolution is not absolutely necessary if the instrumental line shape is accurately known and background effects are not overwhelming, since, in the limit where counting statistics are negligible, a complex spectrum containing a finite number of discrete lines can in principle be analysed uniquely to obtain the component radiations. However, uncertainties due to counting statistics, background, or line shape often cannot be avoided and spectrometers with good resolution are essential for accurate measurement. For example, in the study of γ-ray internal conversion lines superimposed on a β-spectrum continuum, high resolution is invaluable for increasing the peak-to-background counting ratio. In this report we shall, therefore, confine ourselves largely to discussion of instruments giving high resolution. The leading instruments of each type have remarkably similar performance figures, and in general, we shall confine detailed discussion to those which are capable of an energy resolution exceeding 1% and an energy accuracy to better than 0·2%.

It should perhaps be emphasized at the outset that there are several reasons why it is important to measure γ-ray energies to better than 0·2%. Firstly, precise values provide direct determinations of nuclear level energies and binding energies to this accuracy. Theories of nuclear structure (e.g. the collective model, Bohr and Mottelson 1955) demand that level energies be measured with great accuracy in order to provide adequate tests for the theory, particularly when small perturbations are considered (e.g. Kerman 1956), and calculations of nuclear binding energies (Talmi and Thieberger 1956, Goldstein and Talmi 1957) have shown that, at least for some lighter elements, these quantities can be calculated to within a few tens of kilovolts. Secondly, precise γ-ray energies provide a means of discriminating between possible alternative assignments of γ-rays in a decay scheme. For example, in the very complex decay schemes encountered in the (n, γ) reaction in heavy elements the requirement for high accuracy and good resolution for solving the decay scheme would appear to be almost unlimited. Thirdly, of equal importance (since a kindred goal is served in the end) is the necessity of establishing accurately known reference energies for standardization of other instruments or for comparison with the results of independent techniques for absolute energy measurement. Similarly, precise intensity measurements are required for comparison with theory (e.g. Alaga, Alder, Bohr and Mottelson 1955)

and are also indispensable in the procedure of solving a γ-ray decay scheme. For a further discussion of the reasons for precision measurements and for a review of methods applicable to measurement of γ-ray following β-decay the reader is referred to Hedgran (1951).

Four different interactions of γ-rays with matter are commonly used for measurement of γ-ray energies and intensities. These are coherent scattering (crystal diffraction), the photoelectric effect, the Compton effect, and pair production. It would be inappropriate here to enter into a detailed discussion of these various interactions; the reader is referred to the many excellent reviews on this subject (for example, Bethe and Ashkin 1952 or Davisson and Evans 1952). Theoretical expressions describing these processes are introduced as required in the sections dealing with various types of spectrometer. Both the photoelectric effect and pair production have internal conversion analogues. Although, strictly speaking, γ-radiation plays no part in these processes a discussion of γ-ray spectrometers would be incomplete if instruments making use of these effects were ignored. In cases where both external and internal conversion can be used the internal process often gives superior results. Another process sometimes used for γ-ray detection is the photonuclear effect using deuterium. This process has a lower cross section than those for the electronic processes and has not so far been employed for spectrometers giving high precision.

It is not intended that this report present an exhaustive review of γ-ray spectrometers. Rather, an attempt is made to select for detailed discussion those instruments or methods which have given the highest precision in each of the categories defined by the four interactions. Crystal diffraction spectrometers are discussed in § 2, spectrometers using the photo effect in § 3, spectrometers using the Compton effect in § 4, and spectrometers using the pair effect in § 5.

In each section the performances of different types of instrument are compared wherever possible and the applications for which each seems best suited are considered. The discussion will tend to be more detailed for spectrometers with which the authors have had some experience either experimentally or in theoretical design. For other spectrometers the details of design and performance will be drawn from the published literature. The spectrometers which are discussed in some detail are listed in table 1.1 wherein typical performance figures are shown for each instrument with a reference to the published literature.

None of the present authors has had laboratory experience with spectrometers using either the photoelectric effect or Compton effect. For this reason these instruments receive a more superficial treatment than they deserve and the reader is referred to the literature for more detailed discussion. Fortunately, much has been written on the techniques needed in using β-spectrometers for the study of conversion lines, see, for example, the article by Hedgran in *Manne Siegbahn* edited by Melander *et al.* (Hedgran 1951), and several chapters of *Beta and Gamma Ray Spectroscopy* edited by K. Siegbahn (1955). In § 4, we attempt to give a review of the development of the Compton instrument, something which does not appear to have been done before. Such a review is particularly appropriate since most of the work with this type of spectrometer has been done by Russian physicists, many of whose papers have not always been readily available to English-speaking readers.

Table 1.1. Spectrometers with High Precision

Interaction	Type	Typical energy resolution	Useful energy range	Energy precision	Reference
	2m bent crystal line source	1% at 500 kev 0·2% at 100 kev	<1·5 Mev	0·012% at 500 kev 0·03% at 100 kev	DuMond (1947) Beckman *et al.* (1958)
Coherent scattering	7·7m bent crystal line source	0·3% at 300 kev	<1·4 Mev	<0·1% at 300 kev	Rose *et al.* (1957)
	Flat crystal, broad source	0·5% at 1–3 Mev	<4 Mev	0·1% at 1–3 Mev	Knowles (1959)
	Cauchois transmission broad source	~0·1% at 60 kev	<1 Mev	0·02% at 60 kev	Chupp *et al.* (1957 a)
Photoelectric effect	Ring-focus uniform field lens	0·19%	20 kev–5 Mev	0·1% at 1 Mev	DuMond (1957)
	$\pi\sqrt{2}$ high aperture	0·015%	<5 Mev	0·0012% at 662 kev	Graham *et al.* (1959, private communication)
Compton effect	Flat type broad radiator	~2·3%	>200 kev	0·4% at 5 Mev	Groshev *et al.* (1956)
Pair effect	Statistical separation intermediate image lens	1%	>2 Mev	0·15% at 6 Mev	Alburger (1956) Rankin *et al.* (1958)
	Flat type uniform field broad radiator	1%	>2·5 Mev	~0·1% at 8 Mev	Kinsey & Bartholomew (1953)

The suitability of a spectrometer for a particular application depends on many considerations among which are the energy range to be covered, the cross section of the interaction utilized by the spectrometer, the intensity of background radiation from competing reactions and from external sources, and the dimensions of the source. Roughly speaking, the useful range of photoelectric and crystal diffraction spectrometers extends from a few tens of kev up to about 4 MeV, the range of pair spectrometers from about 2·5 MeV upward, while Compton spectrometers can be used from a few hundred kilovolts to 12 MeV and probably much higher. Spectrometers which employ coincidence detection, such as pair spectrometers and Compton spectrometers, are better suited for applications where the intensity of background radiation is high than are crystal diffraction and photoelectric spectrometers which use only a single detector. Spectrometers which are designed to detect divergent radiation from a point source or a line source are usually not well suited for applications in which the source is broad and the radiation well collimated. Point sources are generally used in the study of γ-rays following β-decay and are also encountered when the reaction studied is produced in the target of a charged particle accelerator while broad sources may be produced by neutron bombardment of large targets in a reactor. In § 6, an attempt is made to compare the performance of spectrometers of different types for the measurement of γ-rays from a broad source in a reactor and from a point target in an accelerator.

For the precise measurement of energies with a spectrometer an accurate knowledge of the energy calibration of the instrument is required. In essence, this involves knowing the constants A in an equation for the energy:

$$E_\gamma = A_0 + A_1 x + A_2 x^2 + \ldots \qquad \ldots\ldots(1.1)$$

where x is some instrumental variable such as a magnetic field strength, or an angle. The constants may be determined either absolutely, by direct calculation from the known fundamental parameters of the instrument, or relatively, by calibration using γ-rays of known energy. For many instruments the constants A_1, A_2, etc., may be determined with great precision absolutely while A_0 remains uncertain, that is, the instruments require external calibration for the determination of A_0. However, for the purpose of assigning γ-rays in a decay scheme, an accurate value of A_0 is not required since only the differences between γ-ray energies are important. Thus, in a given experimental field, no spectrometers except those that are capable of making the most accurate absolute measurements need be operated as absolute instruments. The others may be calibrated by comparison with the more accurate instruments or with some other energy standard. In this report, an attempt will be made to distinguish between the performance of various instruments for absolute and relative energy determinations.

In studying the performances of spectrometers of different types it will be of interest to examine the various stages involved in the observation of the γ-ray and to determine the transmission losses suffered at each stage. For spectrometers which depend upon the momentum analysis of secondary electrons, we may consider that the complete system includes a source, converter, analyser, and finally, a detector. The peak counting rate (i.e. the counting rate at the maximum of the line) in counts per second, C, may be expressed in general as

$$C = SA\Omega\epsilon_c T\epsilon_d \qquad \ldots\ldots(1.2)$$

where S is the specific activity of the source in γ-rays per unit area per second (assumed constant over the source area), A is the effective source area, Ω is the fraction of the sphere subtended at the source by the converter, ϵ_c is the efficiency of the converter in electrons per quantum, T is the transmission of the analyser at the peak, and ϵ_d is the efficiency of the final detector in counts per electron. For the internal conversion processes, the source and converter are combined. For a number of instruments discussed below (table 1.2), the functions of analyser and detector are combined. These instruments all happen to have inferior resolution for which the combining of the two functions is not obviously a fundamental cause.

The stages involved in the observation of a γ-ray with crystal diffraction spectrometers differ fundamentally from those for spectrometers based on the other three interactions since no conversion process occurs. For these instruments,

Table 1.2. Spectrometers with Low Precision

Method	Precision	Resolution	Representative References
Cloud chamber with pairs and recoil electrons	± 400 kev near 6 Mev	~13%	Gaerttner & Pardue (1940)
Cloud chamber with $^2H(\gamma, n)\ ^1H$	~5%	~5%	Richardson & Emo (1938)
High pressure ion-chamber with $^2H(\gamma, n)\ ^1H$	100 kev 3–14 Mev	~20% at 3 Mev	Bishop & Wilson (1955)
Emulsions with $^2H(\gamma, n)\ ^1H$	50 kev at 6·15 Mev	~9% at 6 Mev	Gibson *et al.* (1951)
Single NaI crystal	few kev	6% at 0·661 Mev	Bell (1955) Davis *et al.* (1956)
NaI Compton	few kev	7·7% at 1 Mev	Howland *et al.* (1956)
NaI pair	30 kev at 3·86 Mev	6% at 4·43 Mev	Bent & Kruse (1957)
γ-ray hodoscope	1–3%	~5% at 2·75 Mev	Dzhelepov (1957) Chubinskii (1957)

the factors ϵ_c and T in equation (1.2) must be replaced by factors containing the efficiency of the baffle system and the reflection efficiency of the diffracting crystal or crystals.

Methods of γ-ray measurement giving highest precision seem to have two properties in common : (*a*) they have a low efficiency imposed by the converter–analyser system, (*b*) the momentum or energy analysis does not depend on ionization processes.

The less precise methods violate one or both of these ' requirements '. Among such imprecise methods we include cloud chambers, ionization chambers, emulsions, and scintillation methods. Table 1.2 contains a selection of typically good performance figures for instruments based on these latter methods. The reference list is not complete and superior results may have been obtained by other authors who are not included. The γ-ray hodoscope is perhaps out of place in this table. It is a rather novel instrument and calls for some further discussion. It employs

a broad radiator and three arrays of ten, thin-walled, Geiger counters arranged radially at 90°, 180° and 270° from the converter of a flat magnetic field spectrometer. The locations of any three counters, one from each array, which fire in coincidence, define the angle of emission and radius of curvature of the trajectory of a secondary electron ejected from the converter. The device has the virtue of having an extremely low background and may be used to detect simultaneously γ-rays of different energies. The inclusion of this method in the table is perhaps justifiable because the three counters introduce a straggle in the electron trajectory which is not present in other methods of magnetic analysis and because the resolution is limited by the size of the counters. Beyond this brief reference, the methods described in table 1.2 will receive no further attention in this report. Nor shall we discuss in detail any novel methods which show promise for precision measurement but which have not yet been perfected in practice. Among such methods, we draw attention here to the trochoidal orbit spectrometer being developed by Malfors (1958). This instrument is expected to have high energy resolution and high transmission, and also may be arranged to record all radiations in an extended spectrum simultaneously.

§ 2. Spectrometers using Coherent Scattering

2.1. *Introduction*

This section is concerned with the diffraction of gamma radiation from single crystals and the determination of energies and intensities of the incident γ-rays from the angular distribution of the diffracted radiation.

In discussing the diffraction of radiation by a crystal, we consider only that radiation incident on the crystal which is scattered elastically, without loss of energy, by the spatial (time average) distribution of electrons surrounding each atom. The process is completely analogous to diffraction of light by a grating since the electron distribution in a crystal appears, to the incident γ-ray, as a three-dimensional distribution with periodic scattering power. Diffraction maxima are expected in directions corresponding to path differences of integral number of wavelengths. What these directions are, in terms of the direction of the incident γ-ray relative to the orientation of the crystal, is shown (the analogy of W. H. and W. L. Bragg) by comparing the diffraction of radiation from a crystal with the reflection of light from a mirror. The periodic distribution of atoms in a crystal is described in terms of a set of equally spaced parallel planes. The orientation of the planes, usually with respect to three simple axes of the crystal, is defined by a normal vector $\mathbf{B}_H$. For a cubic ' type ' lattice the subscript H represents the three small integers h_1, h_2, h_3 (Miller indices) which are proportional to the direction cosines of the normal to the set of planes. The simple crystal axes are usually non-orthogonal and assigned different unit lengths. In general, the indices h_1, h_2, h_3 are defined as a set of small integers which are proportional to the reciprocal intercepts of the parallel crystal planes on the three chosen axes, the intercepts being measured in their respective axial units. A plane wave of electromagnetic radiation, diffracting from a single atomic plane, is completely in phase after diffraction if the angle of incidence equals the angle of diffraction, analogous to the reflection of light from a plane mirror. But the diffracted waves from different planes of the set, spaced

d_H apart, are only in phase for particular angles of incidence, $90° - \theta_B$, where θ_B, the complement of the angle of incidence, is determined by the Bragg equation

$$n\lambda_x = 2d_H \sin\theta_B \qquad \ldots\ldots(2.1)$$

where λ_x is the wavelength of the radiation, d_H is the plane spacing, and n is the order of the reflection.†

It is inferred that θ_B in equation (2.1) applies to that scattering angle which corresponds to the maximum intensity of the diffracted line. Although this interpretation would be correct on the basis of the simple analogies used to describe the diffraction process, it is not necessarily correct for diffraction in real crystals in which refraction and absorption as well as diffraction occur. For example, the wavelength and direction of the radiation may undergo a slight change due to refraction on passing through the crystal surface. Equation (2.1) applies to the wavelength and angle of incidence of the radiation inside the crystal. However, if the γ-rays are transmitted by a crystal slab with its diffraction planes perpendicular to the incident and exit faces, the change in wavelength and direction at the crystal surfaces is such that equation (2.1) applies to both the internal and external wavelength and the direction of radiation. Nearly all crystal spectrometers considered in §§ 2.2 and 2.3 use crystals which are cut in this manner.‡ Equation (2.1) correctly relates the γ-ray energies and the diffraction angle measurements made with these spectrometers to better than one part in 10^5.

In γ-ray spectrometry it is the energy of the radiation E_γ rather than the wavelength λ_x which is of interest. These are related by the equation

$$E_\gamma \lambda_x = hc = 12{\cdot}39767 \pm 0{\cdot}00022 \text{ MeV MÅ} \S \qquad \ldots\ldots(2.2)$$

where h is Planck's constant and c is the velocity of light. λ_x is the wavelength of the radiation in absolute MÅ units and E_γ the energy in MeV. In many incidences only relative γ-ray energies are required. For such cases equations (2.1) and (2.2) and the Bragg diffraction angles for two γ-rays are sufficient to determine the ratio of their energies. For the absolute determination of γ-ray energies, a knowledge of the plane spacing d_H is necessary. The spacing d_H can be obtained by measurement using X-rays or γ-rays of known wavelengths.‖

Now we shall consider a diagrammatic spectrometer, figures 2.1 (*a*) and (*b*), designed to measure the direction of radiation diffracted from a crystal. This spectrometer consists of a source S, a baffle B (this baffle is omitted in figure 2.1 (*a*))¶ and a detector D. Figure 2.1 (*a*) illustrates the three-dimensional aspects of the spectrometer : in particular, it shows the radiation paths from source

† n is an integer which is the number of wavelengths in the radiation path difference between successive diffracting planes.

‡ Similarly, effects of γ-ray absorption in near perfect crystals influence the symmetry of the diffracted line but disappear when the crystal surfaces are perpendicular to the diffracting planes (Zachariasen 1946).

§ The value of $\lambda_x E_\gamma$ has been estimated by Cohen *et al.* (1955) by appropriate combination of different precise measurements of the atomic constants.

‖ Since X-ray wavelengths are quoted in X units d_H measurements which refer to them must be multiplied by a factor $1{\cdot}002037 \pm 0{\cdot}000020$ (DuMond 1959, see also Merrill and DuMond 1958) to convert from X units to MÅ units. However, the determination of d_H by calibration against a γ-ray of known energy is made entirely in absolute units.

¶ An alternative geometry in figure 2.1 (*b*) would be to define the direction of the incident radiation by the width Q′ Q″ of the diffracting plane bounded by the crystal faces CC, rather than by the width of the aperture in baffle BB, but this would have complicated the discussion.

to crystal which deviate from the ZX plane. Radiation emitted from a point S in the source and incident on the crystal Q at an angle $90° - \theta_B$ generates a segment of a cone with axis SO parallel to $\mathbf{B}_H$ the normal to the diffracting planes. The energy E_γ and incident angle $90° - \theta_B$ of this radiation satisfy the Bragg equation (2.1) so that radiation is diffracted with maximum intensity into the detector D.

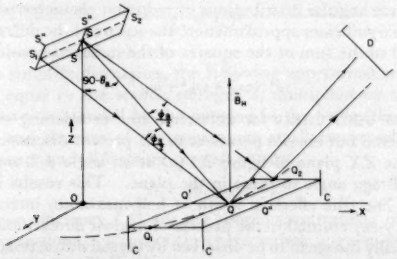


Figure 2.1 (*a*). A diagrammatic spectrometer consisting of a source S, a crystal C, and a detector D, shows the rays incident on the crystal at angles $\phi_S/2$ and $\phi_B/2$ to the ZX plane. $\mathbf{B}_H$ is the normal to the diffracting planes of the crystal.

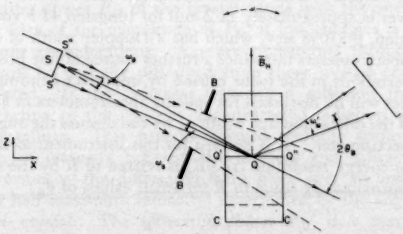


Figure 2.1 (*b*). A view of the same spectrometer in the ZX plane, except that a baffle B has been added, shows the characteristic widths ω_S, ω_B and $\omega_{M'}$.

The length of the source S_1S_2 and the length of the crystal Q_1Q_2 (parallel to the Y axis) are always small compared with the distance SQ between them so that γ-rays incident on the crystal are no more than a few degrees (angles $\phi_S/2$ and $\phi_B/2$) out of the ZX plane. Since the Bragg angle θ_B is usually small (it is 3° for the diffraction of 100 kev γ-rays from the (310) planes of quartz) the diffraction cone Q_1SQ_2 is nearly flat and cuts the atomic planes of the crystal at Q_1QQ_2 in an almost straight line. Under these conditions radiation emitted from all points of the source and diffracted by the crystal may be treated with sufficient accuracy by considering it to be parallel to the ZX plane, as illustrated in figure 2.1 (*b*). Marked in figure 2.1 (*b*) are the angle ω_S subtended at the crystal by the width of the source, the angle ω_B subtended at the crystal by the width of the baffle aperture and the angle

ω_M', the width at half-maximum intensity of a parallel monochromatic plane wave after diffraction by the crystal.†

If the crystal is rotated about a line through Q parallel to the Y axis (note that the broad detector D is fixed) the diffracted intensity will pass through a maximum value near the angular position shown in figure 2.1 (*b*). The angular width ω of this diffracted line at half-maximum intensity is in general a complicated function of the three angular distributions of radiation characterized by the widths ω_S, ω_B and ω_M'. To sufficient approximation, the square of the diffraction line width ω may be equated to the sum of the squares of the individual widths :

$$\omega^2 \simeq \omega_S^2 + \omega_B^2 + \omega_M'^2. \qquad \ldots\ldots(2.3)$$

There are three other causes for diffraction line broadening which should, in general, be considered but are not pertinent to the present discussion. The γ-rays directed out of the ZX plane of figure 2.1 (*a*) at an angle $\phi/2$ are diffracted at a slightly different Bragg angle to those in the plane. This results in a broadening of the diffraction line (the effective width at half-maximum intensity $\omega_\phi \simeq \phi^2\,\theta_B$). Also, a ' discrete ' γ-ray emitted in the process of nuclear de-excitation has a natural width which is usually too small to be observed by crystal diffraction. For example, a 200 kev γ-ray which has an energy width at half-maximum intensity of 10^{-2} volt will produce a diffracted line width of only 10^{-4} second of arc when diffracted from the (211) planes of calcite. However, appreciable broadening may be observed for x-rays which have natural widths at half-maximum intensity of some volts (the K_α doublet for silver is approximately 11·2 and for tungsten 42·5 volts) and for the annihilation radiation, 0·51098 Mev, which has a Doppler width of a few hundred volts. Focusing spectrometers introduce a further source of line broadening which is attributed to aberrations in the focus caused by imperfect bending of the crystal lamina. This effect will be discussed for specific spectrometers in § 2.2.

Having defined the diffraction line width ω, we can discuss the angular resolution R of the simple spectrometer. It is defined for this instrument as the ratio ω/θ_B.‡ The corresponding energy resolution $\Delta E_\gamma/E_\gamma$ is related to R by the equations (2.1) and (2.2) and is approximately equal to R for small values of θ_B :

$$R = \Delta E_\gamma/E_\gamma \simeq \omega/\theta_B. \qquad \ldots\ldots(2.4)$$

The transmission efficiency I_p/I_s of a spectrometer is defined as the ratio of the maximum intensity (peak intensity) I_p of the diffracted line to the effective source strength I_s. To sufficient accuracy the peak intensity I_p the area under the diffracted line I_A and the observed line width ω are related by the approximation

$$I_p \simeq I_A/\omega. \qquad \ldots\ldots(2.5\,a)$$

† ω_M' is one to two seconds of arc for nearly perfect crystals of quartz and calcite. For γ-rays with energies greater than about 40 kev this width is determined by, and is nearly equal to, ω_M the angular distribution, called the mosaic, of the crystallites which constitute the crystal. Diffraction broadening, the width resulting from diffraction by a limited number of atomic planes, is negligible for these energies. It is inversely proportional to the energy of the diffracted radiation and becomes appreciable below 20 kev. A detailed discussion of the perfect crystal diffraction line width is given by Zachariasen (1946).

‡ This definition for R is used for all single crystal instruments, but in order to include spectrometers which use more than one crystal, the resolution is defined more generally by $R = \Delta E_\gamma/E_\gamma \simeq {}^T\omega/\theta_T$ where ${}^T\omega$ is the total observed diffraction line width at half-maximum intensity and $2\theta_T$ is the total scattering angle of the incident radiation.

In order to express the ratio I_p/I_s in terms of characteristic parameters of the simple spectrometer, it is necessary to define the reflecting power P_H of the diffracting crystal. If I_0 is the intensity in counts per second of radiation of energy E_γ transmitted through the baffle, figure 2.1 (*b*), from the source at S, and I_A is the area under the diffracted line in count radians per second obtained by rotating the crystal through the Bragg angle, the reflecting power P_H, in radians, of the crystal is defined by the equation

$$I_A = I_0 P_H. \qquad \ldots\ldots(2.5\,b)$$

P_H will be discussed in some detail.

In order to simplify discussion, the following approximations are made: I_0 is approximately equal to the source strength I_s multiplied by the fractional solid angle at the source subtended by the opening in the baffle (the dimensions of the baffle opening and the size of the source are small compared with the distance between the source and diffracting crystal) so that

$$I_0 \simeq (\phi_B \omega_B/4\pi) I_s.$$

The transmission efficiency I_p/I_s for the simple spectrometer obtained from these approximate expressions in terms of the characteristic angular widths ϕ_B, ω_B, ω and the crystal reflecting power P_H is

$$\frac{I_p}{I_s} \simeq \frac{\phi_B \omega_B}{4\pi} \frac{P_H}{\omega}. \dagger \qquad \ldots\ldots(2.6)$$

The reflecting power P_H of the crystal varies with the energy of the incident γ-ray. Some idea of the magnitude of P_H at low energies is obtained from the following simple considerations. A γ-ray transmitted through a crystal at the Bragg angle will be partially diffracted from those mosaic crystallites which are properly oriented. If the radiation traverses a sufficient thickness of crystal, repeated diffraction of the incident and diffracted radiation will occur from the individual crystallites. This process, termed multiple scattering, results in an equal division of intensity between the incident and diffracted radiations. As mentioned previously, the diffraction takes place over an angular range (with a width ω_M'‡ at half-maximum intensity) determined by the angular distribution of mosaic in the crystal. The reflecting power P_H' is a maximum under these conditions and is expressed in terms of ω_M' and the crystal transmission absorption factor e^{-ut}, by the approximate equation

$$P_H' \simeq \tfrac{1}{2}\omega_M' \exp(-ut) \qquad \ldots\ldots(2.7)$$

where t is the path length and u is the linear absorption coefficient of the radiation. Equation (2.7) would apply to a 0·5 MeV γ-ray diffracted from a near perfect crystal, such as a 1 cm thick calcite, which has a mosaic distribution width $\omega_M \simeq 2$ second. In this case (neglecting γ-ray absorption in the crystal) $P_H' \simeq 10^{-5}$ radian. If we use this calcite crystal in the simple γ-ray spectrometer of figure 2.1 (*b*) (the

† For many spectrometers ω_B, ω_s and ω_M' and therefore ω are constant. When this is the case the ratio P_H/ω, called the crystal reflectivity, is a useful quantity describing the spectrometer. In general also, when ω_s and ω_B are small compared with ω_M', P_H/ω is dependent only on the properties of the diffracting crystal.

‡ When there is pronounced secondary extinction in the crystal ω_M' is not equal to the width ω_M of the mosaic distribution of the crystal but is usually somewhat larger. Secondary extinction is the depletion of the incident and diffracted beams by diffraction in the crystal.

efficiency of which can be estimated from equation (2.6)) and assume for simplicity that the direction of the incident radiation is only roughly defined (which is so when $\omega \simeq \omega_B \gg \omega_M$ and $\phi_S \simeq \phi_B \simeq 1$ degree) then the efficiency will be approximately 10^{-8}. The spectrometer efficiency is even less for diffraction of higher energy γ-rays because the reflecting power of the crystal decreases with increasing energy. At sufficiently high energies, multiple scattering and secondary extinction of the radiation transmitted through the crystal become negligible. Under such conditions the crystal reflecting power is proportional to the average number of mosaic crystallites in the path of the transmitted beam (this number is roughly t/t_s where t is the radiation path length in the crystal and t_s the average thickness of a mosaic crystallite) times the average reflecting power of a crystallite. Since a crystallite is assumed to be perfect, its reflecting power is proportional to the average angular spread of the diffracted radiation times the square of the sum of the amplitudes of the radiation scattered from the crystal unit cells. To sufficient approximation, the interference width at half-maximum intensity is approximately equal to d_H/t_s and the scattered amplitude to $r_0(F_H/v)\lambda_\gamma t_s$ where r_0 is the electron radius, F_H is the crystal structure factor and v is the unit cell volume. The reflecting power R_H for a high energy γ-ray diffracted, without absorption, during transmission through the crystal is obtained from the above factors and is given approximately by the equation

$$R_H \simeq r_0^2(F_H/v)^2 \lambda_\gamma^2 d_H t. \qquad \ldots\ldots(2.8)$$

When the absorption of radiation in the crystal is included the reflecting power P_H'' is

$$P_H'' \simeq R_H \exp(-ut). \qquad \ldots\ldots(2.9)$$

These considerations, summarized by equation (2.8), show that the reflecting power of a crystal for very high energy γ-rays decreases rapidly with decreasing wavelength, or increasing energy, (as λ^2 or E_γ^{-2}). In figure 2.2 are plotted reflecting powers P_H† as a function of γ-ray energy for some of the crystals used in the spectrometers discussed in §§2.2 and 2.3.

The simple spectrometer described has a very low transmission efficiency when the incident radiation is confined (by a small source and baffle width) to the narrow beam necessary for good resolution. There are two ways that this spectrometer can be modified (or added to) in order to improve the counting rate at the detector.

The peak counting rate may be improved by a factor N, with little or no loss in resolution, by simply increasing the width of source and diffracting crystal by a factor N. N parallel channels (called Soller slits) are used to define the incident direction of the radiation with respect to the crystal, figure 2.3 (*a*). This is the broad source, single crystal transmission spectrometer described in §2.3. The

† The values of P_H have been calculated from the equation

$$P_H \exp(ut) = \tfrac{1}{2}\int \{1-\exp[-2W(\theta-\theta_B)\,R_H]\}\,d\theta \qquad \ldots\ldots(2.10)$$

where the integration extends over the angular spread of the diffracted line. This expression was derived by Knowles (1959), assuming that there was no primary extinction of the transmitted beam in the crystal but only secondary extinction and multiple scattering. In this expression $W(\theta-\theta_B)$ is the angular distribution of the mosaic blocks. This distribution is assumed to have a Gaussian form in the calculations of P_H in figure 2.2. Knowles showed that when $R_H\,W(\theta-\theta_B) \ll 1$ (the condition for a poor reflecting crystal) the integral for P_H reduces to equation (2.9) and when $R_H\,W(\theta-\theta_B) \gg 1$ (the condition for a strong reflecting crystal) the integral can be approximated by equation (2.7).

peak intensity I_{pS} and transmission efficiency I_{pS}/NI_s, obtained from equation (2.6) in terms of the parameters for the simple spectrometer are given by the equation

$$\frac{I_{pS}}{NI_s} \simeq \frac{\phi_B}{4\pi}\frac{\omega_B}{\omega} P_H \qquad \ldots\ldots(2.11)$$

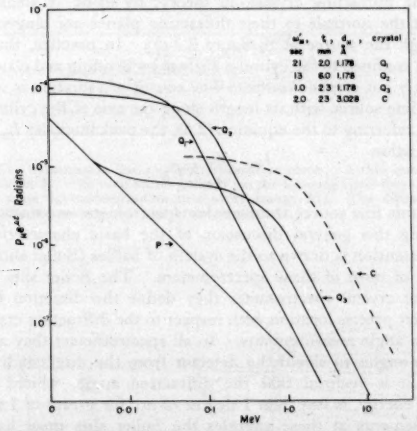


Figure 2.2. The calculated reflecting power of the (310) planes of quartz, $d_{310} = 1{\cdot}178$ Å, divided by $e^{-\mu t}$ the γ-ray absorption in the crystal, is plotted as a function of the energy, in MeV, of the diffracted radiation for three thicknesses of quartz crystal: Q_1, 2 mm, Q_2, 6 mm, Q_3, 23 mm. Crystal thicknesses t, mosaic distribution widths ω_B' and reflecting plane spacings d_H correspond to crystals used in spectrometers discussed in the text. The straight line P gives the reflecting power for a perfect quartz crystal. At sufficiently low energies the diffraction efficiency of quartz is independent of crystal thickness, but at higher energies (depending on the crystal perfection) the efficiency is greater than that given by line P. The broken line C is the calculated reflecting power of the (211) planes of a calcite crystal 23 mm thick. This crystal is used in the flat crystal spectrometer.

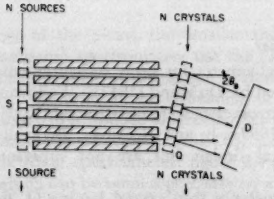


Figure 2.3 (*a*). A diagrammatic sketch of a broad source single flat crystal spectrometer illustrating its similarity to the simple diagrammatic spectrometer of figure 2.1.

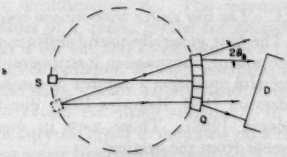


Figure 2.3 (*b*). A diagrammatic sketch of a line source spectrometer employing individually oriented flat crystal blocks which simulate a bent crystal.

where $NI_s = I_s'$ is the total source strength. A practical variation of this broad source spectrometer is the double flat crystal transmission spectrometer. Both of these spectrometers are described in detail in §2.3.

A second method of improving the counting rate of the simple spectrometer by a factor N is to increase the transmission efficiency. This is done by increasing the width of the diffracting crystal, in theory, by using N small crystals, each oriented so that the normals to their diffracting planes are tangent to a cylinder with axis through the source at S, figure 2.3 (*b*). In practice, the crystal planes are oriented and confined to the cylinder surface by bending and clamping on to this surface a properly cut crystal lamina. The source of radiation is not increased in size (it is a thin line source with its length along the axis of the cylinder). For this geometry, again referring to the equation (2.6), the peak intensity I_{pB} is given by the approximate equation

$$\frac{I_{pB}}{I_s} \simeq \frac{\phi_B}{4\pi} N P_H. \qquad \ldots\ldots(2.12)$$

Other forms of this line source transmission spectrometer are discussed in §2.2.

In concluding this general discussion of the basic characteristics of crystal spectrometers, attention is drawn to the system of baffles (Soller slits) which are an essential feature of most of these spectrometers. The Soller slits have two uses. In the single flat crystal spectrometer they define the direction of the incident radiation and their precise location with respect to the diffracting crystal is essential to the diffraction angle measurements. In all spectrometers they are necessary at small diffraction angles to shield the detector from the direct radiation. This is apparent when it is realized that the diffraction angle, which decreases with increasing γ-ray energy, is less than 1 degree of arc for γ-rays of 1 MeV. In order to make measurements at these energies the Soller slits must have an angular aperture of only a few minutes of arc.

2.2. *Line Source Spectrometers*

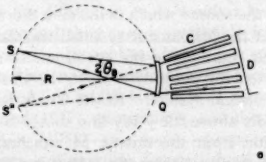
The line source transmission spectrometer is the more generally used and will be considered in detail. The reflection type has been used for the study of low energy radiation, x-rays, but has not been used for high energy γ-ray measurements. However, the line source reflection spectrometer has characteristics which would be useful in a high energy instrument and it will be discussed briefly at the conclusion of this section.

2.2.1. *The line source transmission spectrometer.*

The first generally useful diffraction spectrometer for γ-ray measurement was a line source transmission instrument designed by DuMond (1947). Following its successful operation a number of similar spectrometers were constructed (Ryde and Andersson 1955, Knowles 1957, Sumbaev 1957 a, b, and Beckman, Bergvall and Axelsson 1958). These vary in mechanical design but are not substantially different from the original.

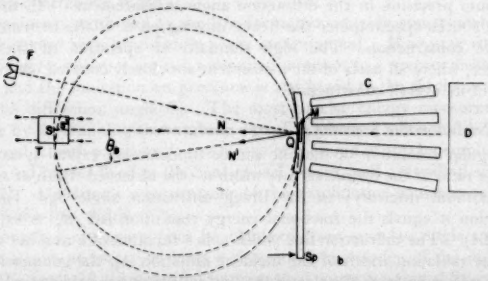
In the line source transmission spectrometer a thin crystal lamina Q, figure 2.4 (*a*), is bent to such a curvature that γ-rays emitted from a line source S on the focusing circle are diffracted from the crystal planes and detected by an extended detector D. Each position of the source on the focusing circle corresponds to a

particular energy, so that motion of the source along this circle allows the γ-ray spectrum to be measured. A particular γ-ray emitted from a source at S is diffracted from the crystal planes if its energy and angle of incidence satisfy the Bragg equation (2.1). As expressed by equations (2.1) and (2.2), the greater the



a

Figure 2.4 (*a*). The transmission line source crystal spectrometer. A thin quartz crystal lamina Q is bent to radius *R*. For each source point S′ on the focusing circle there corresponds a Bragg diffraction angle θ_B (corresponding to a γ-ray energy E_γ). The diffracted beam proceeds through Soller slit system C to the broad detector D.



b

Figure 2.4 (*b*). The Chalk River version of the DuMond 2-metre spectrometer. The γ-ray and light sources are at S. They are translated to S′ on the trolley T. The precision table Sp rotates crystal Q from angular position N to N′ by means of a tangent screw. The spectrometer automatically rotates (following the rotation of turntable Sp) by means of an optical system which includes the mirror M and photocell F such that the diffracted γ-ray proceeds down the stationary Soller slit C to the detector D.

energy of the γ-ray, the smaller the angle of diffraction θ_B. The majority of transmission spectrometers use the (310) planes of a quartz crystal for diffraction so that the angle of diffraction is about 3° for a 100 kev and 18 minutes for a 1 Mev γ-ray. Because of the proximity of the direct radiation at high energies, it is necessary to introduce tapered baffles C (Soller slits), usually made of lead, in order to shield the detector (a NaI (Tl) scintillator) from the direct radiation. In the Chalk River spectrometer, figure 2.4 (*b*), the curved crystal is mounted on a table Sp which can be rotated precisely by a tangent screw from angular position N to N′. The imaginary focusing circle rotates with the crystal until the Bragg equation (2.1) is satisfied for a particular γ-ray. Following this rotation the source S is slightly off the focusing circle. To preserve the resolution, the source, mounted on a trolley T, is moved forward to the approximate position S′. A further rough adjustment

is necessary, the orienting of the slits and detector, so that the diffracted γ-ray beam will proceed through the Soller slit C. With the DuMond and Chalk River spectrometers this is realized by leaving the slits fixed and rotating the whole spectrometer and source about a bearing at Q. With the Argonne 7·7 metre spectrometer (Rose *et al.* 1957) the bent quartz crystal and slit system can be rotated while the motion of the source which is inside a thermal neutron reactor is restricted within the focus of the instrument to small translations in the direction of the bent crystal. Angular rotations of the spectrometer relative to the axis of the Soller slit are controlled to a few seconds of arc in the Chalk River and Argonne spectrometers by use of an optical system. In the Chalk River spectrometer the light source is mounted rigidly above the γ-ray line source S′, figure 2.4 (*b*), and a parallel beam, after reflection from the mirror M attached to the bent crystal turntable, is located by a detector consisting of a split prism with photocells F fixed in the laboratory. Mechanical decoupling of the light weight precision mechanical movements (the rotation of the crystal turntable) from the less precise heavy mechanical motions (the spectrometer rotation relative to the Soller slit) ensures the maximum precision in the diffraction angle measurements. In the Knowles (1957) line source spectrometer the heavy moving parts of the instrument are of quite crude construction. The high standard of operation of the DuMond spectrometer, where all parts of the instrument are closely coupled, stems from the excellence of its overall construction.

2.2.1.1. *Resolution and precision of energy measurements.*

The angular resolution of the line source transmission crystal spectrometer is equal to the ratio of the diffracted line width ω (the observed width of a γ-ray line at half-maximum intensity) to the Bragg diffraction angle θ_B. To sufficient approximation it equals the fractional energy resolution $\Delta E_\gamma/E_\gamma$ as expressed by equation (2.4). The diffraction line width ω is a function of a number of separate widths : the radiation width of the incident radiation ω_γ, the geometrical widths ω_S and ω_B, stemming from the size of the line source and the thickness of the bent crystal respectively, the width ω_I resulting from the aberrations in the focus of the bent crystal, and the crystal diffraction width ω_M'. The observed width may be expressed in terms of the separate widths, see equation (2.3). The width ω_γ is usually negligible for γ-rays. The diffraction width ω_M' is independent of the energy of the γ-ray for the thin crystal laminae used in these spectrometers. The line source width ω_S is constant for each source and may be made negligible by making the source suitably thin. In any case the observed line width ω of bent crystal spectrometers is a constant for each experiment, independent of γ-ray energy, so that the resolution R is inversely proportional to the Bragg angle θ_B, or directly proportional to the γ-ray energy E_γ. A bent quartz crystal spectrometer, in which the crystal has a two-metre radius of curvature, and diffraction is from the (310) planes, has a resolution of about 0·2% at 100 kev and 2% at 1 Mev. The Argonne 7·7 metre bent crystal spectrometer has a factor 1·7 better resolution than this because of its proportionally narrower diffraction line width ω.

The resolution of the line focus spectrometer may be improved (decreased) for a given energy of γ-ray either by increasing the Bragg angle θ_B or by decreasing the line width ω. Most quartz crystal spectrometers diffract from the (310) plane,

$d_{310} = 1{\cdot}178$ Å, which corresponds to a diffraction angle $\theta_B = 1°$ for a 300 kev γ-ray. In order to increase this angle the γ-ray may be diffracted from a crystal plane of smaller spacing. In quartz such planes have poor reflecting power and have been used only for the diffraction of intense γ-rays (Muller *et al.* 1952, Rose *et al.* 1957). The line width ω may be reduced by making the geometric widths ω_S and ω_B negligibly small (since the diffraction angle is only a few degrees of arc the effective width ω_B due to the crystal thickness is negligible for crystals a few millimetres thick). A monochromatic γ-ray emitted from a line source, less than 0·01 cm wide and a few millimetres deep, and incident on a bent quartz crystal 1 to 2 mm thick will result in a diffracted line width less than 21 seconds of arc at half-maximum intensity in the DuMond (1947) spectrometer, and about 15 seconds in the Sumbaev (1957 a) and Beckman (1958) two-metre spectrometers. With the Argonne 7·7 metre spectrometer a diffraction line width of 13 seconds of arc has been attained using a bent quartz crystal 6 mm thick (Rose *et al.* 1957). The minimum line widths reported by these workers were provisionally attributed by them to the bent crystal diffraction width ω_M'† and were thought to be nearly independent of the focusing aberrations width ω_f.

The resolution attributed to the line focus transmission spectrometer does not necessarily ensure high precision in γ-ray measurement. If a γ-ray is sufficiently intense, counting statistics are not a serious limitation to a precise measurement of its energy and the limitation on precision is attributed to the mechanical measurement of the diffraction angle θ_B. The most precise energy measurements have been made by determining this angle to a small fraction of a line width. DuMond has measured it with precision (a few tenths of a second of arc) by employing a light source, curved mirror fixed to the bent crystal mount, and a standard scale (Muller *et al.* 1952). DuMond's measurement of the annihilation radiation was made to about 10^{-2} of a line width.

The correct γ-ray energy can be obtained from such an angle measurement because it is not necessary to have a theoretical knowledge of the diffracted line shape. The angle $2\theta_B$ is measured between two corresponding diffraction lines of a γ-ray which is diffracted by the crystal, to the right and left of the direct beam. These two lines are identical in shape and are not mirror images in the $\theta_B = 0$ plane. Any consistent scheme of defining the line position (preferably one that uses the total count of the diffracted line area) is satisfactory for accurate energy comparison.‡ Muller *et al.* (1952) proposed a method of superposition of the right- and left-hand lines, while Knowles (1959) used the centroid of the lines.

The most precise measurement of low energy x-rays using a two-metre spectrometer has been reported by Beckman *et al.* (1958). They made measurements of the Bragg angle θ_B reliable to $\pm 0{\cdot}3$ second of arc. For the angle measurements they used a precisely divided circle which had previously been used for precision x-ray measurements of the Bragg angle. With this spectrometer Bergvall (1959) compared the K_{α_1} and K_{α_2} x-ray energies of the rare earths to that of MoK_{α_1} to better than 2 parts in 10^5. Table 2.1 is a summary of some of the best γ-ray energy

† The minimum diffraction line width ω_M' exhibited by a perfect elastically bent crystal has been explained by Sumbaev (1957 b) in terms of the elastic deformation of the crystal without postulating an increase in the angular distribution of the crystal mosaic.

‡ This is not necessarily true for γ-rays which have a natural width which is broad compared with the diffracted line width.

Table 2.1

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
DuMond (1947)	22	1·178	0·7	^{193}Pt	308·454 ± 0·033‡	1 : 10^4	0·31
† r.c. 200 cm quartz,			0·9	^{198}Hg	411·770 ± 0·036‡	0·8 : 10^4	0·19
d_{310}			1·2	^{64}Cu	510·941 ± 0·067‡	1·3 : 10^4	0·23
Beckman (1958)	23	1·178	0·20	U K_{α_1}	98·4393 ± 0·0018	3·2 : 10^5	0·35
† r.c. 200 cm quartz,	15		0·57	^{198} Hg	411·770 ± 0·033§	0·8 : 10^4	0·2
d_{310}	27		0·08	La K_{α_2}	33·0340 ± 0·0004‖	1·2 : 10^5	0·3
	25		0·13	Lu K_{α_1}	54·0698 ± 0·0009‖	1·7 : 10^5	0·3
Chupp *et al.* (1957 a, b)	24	1·178		^{159}Tb	57·99 ± 0·01¶	1·7 : 10^4	3·1
† r.c. 200 cm quartz,				^{181}Ta	165·15 ± 0·11¶	6·7 : 10^4	4·3
d_{310}							
Rose *et al.* (1957)	13	1·178	0·21	^{198}Au	158·4 ± 0·1††	6·3 : 10^4	4·0
† r.c. 770 cm quartz,			0·50		381·0 ± 0·4††	1·0 : 10^5	2·6
d_{310}							
Knowles (1959)	3·6	3·028	0·5	^{24}Mg	1368 ± 1	1 : 10^5	0·46
double flat crystal	1·6		0·6	^{24}Mg	2750 ± 3	1 : 10^5	0·30
calcite, d_{211}	3·6		0·5	^{49}Ti	1378 ± 1	1 : 10^5	0·30

(1) Spectrometer description ; (2) line width ω (sec) ; (3) d_H (Å) ; (4) resolution (%) ; (5) isotope ; (6) best measurement (kev) ; (7) best fractional precision ; (8) best angular precision $\pm\Delta\theta$ (sec).

† r.c. = radius of curvature. ‡ Muller *et al.* (1952). § Bergvall (1959, private communication). ‖ Bergvall (1959). ¶ DuMond (1958 b).

†† Hamermesh (1956, private communication), recent measurements are comparable in precision with the 2-metre spectrometer.

measurements made with crystal diffraction spectrometers. The most precise angular measurement, $\pm 0{\cdot}2$ second of arc, has been made by DuMond and independently by Bergvall and corresponds to a precision of 0·8 parts in 10^4 in the energy measurement of 411·770 kev γ-ray of ^{198}Hg. This measurement was made with two-metre bent crystal spectrometers diffracting from the (310) planes of quartz.

For absolute energy measurements the spacing between the crystal reflecting planes d_H must be known with an accuracy comparable with that of the Bragg angle θ_B (see equations (2.1)). This spacing, even in near perfect crystals like quartz and calcite, varies slightly from sample to sample (for calcite this variation is a few parts in 10^5). For this reason precise absolute measurements must include an accurate calibration of the diffracting crystal. DuMond compared γ-ray and x-ray energies by measuring them in the two-metre spectrometer. Relative x-ray energies tabulated in x units† are known to a precision of about 1 part in 10^4 to 10^5, and the uncertainty in the conversion from x units to the absolute MÅ units is estimated as 1·6 parts in 10^5 (DuMond 1959, see also Merrill and DuMond 1958). For this reason there is an uncertainty of 1·6 parts in 10^5 in the determination of absolute γ-ray energies by x-ray comparison.

2.2.1.2. *Efficiency and range of energy measurement.*

The transmission efficiency I_{pB}/I_s for the line source transmission spectrometer in terms of characteristic spectrometer widths, $\bar{\omega}$, ω_B and ϕ_B, and the reflecting power P_H of the diffracting crystal, is given approximately by equation (2.12). Since the diffracting crystal is a single lamina bent to a cylindrical surface, rather than a number of individually oriented crystal blocks, it is better to express the spectrometer efficiency in terms of the effective source strength I_s, the solid angle Ω subtended at the source by the bent crystal and the crystal diffraction efficiency (P_H/I_s). Equation (2.12) is rewritten in the approximate form

$$I_{pB}/I_s \simeq \Omega c'(P_H/\omega). \qquad \ldots\ldots(2.13)$$

c' is the fractional transmission of the system of Soller slits and is about 0·3 to 0·5.

The bent quartz crystal used in line focus transmission spectrometers is only a few millimetres thick and its crystal diffraction line width ω_M' a few seconds of arc. Under these conditions the crystal reflecting power P_H may be calculated from equation (2.9). It is given as a function of γ-ray energy for different crystals in figure 2.2.

Line focus transmission spectrometers have been used to measure γ-ray energies between 8 kev and 1·3 Mev. At low energies, particularly below 40 kev, the source absorption in the line source spectrometer is very great (the effective source strength I_s for a 30 kev γ-ray may be less than one-tenth that of a 100 kev γ-ray of comparable emission strength). However, the reflecting power P_H of a 1 mm thick quartz crystal for such a γ-ray is relatively high (it is even higher for certain reflecting planes in 1 mm crystals of topaz). Very thin crystals (less than 1 mm thick) can be bent to a small radius of curvature, suitable for a spectrometer with

† The present standard for the x unit is the $d_{311} = 3{\cdot}029040$ Å spacing of calcite. When Siegbahn used this value to establish a measuring unit for x-ray wavelengths (which he intended to be one milliångström, 10^{-11} cm) he did not know that this spacing was in error because Avogadro's number, required in the computation, was in error (DuMond 1959, see also Merrill and DuMond 1958).

a large solid angle Ω. Such spectrometers have been constructed by Barton (UCRL report 670, see also Barton, Robinson and Perlman 1951), Browne (UCRL report 1760) and Day (1955) for the study of the L x-rays and the low energy γ-rays of heavy elements. The spectrometer described by Day used a topaz crystal 0·2 mm thick and diffracted γ-rays from the (303) planes, $d_{303} = 1{\cdot}352$ Å. The radius of curvature of the bent crystal was 10 inches and the fractional solid angle Ω greater than 10^{-4}. The transmission efficiency I_{pB}/I_s is $1{\cdot}7 \times 10^{-6}$, for 60 kev γ-rays. The spectrometer has been used to study sources of a few millicuries.

Two-metre curved crystal spectrometers with 1 to 2 mm thick quartz crystals have been used for the study of γ-rays following β-decay between 30 kev and 1·3 Mev. The 7·7 metre curved crystal spectrometer has been used, in the same energy range, for the measurement of the γ-rays following the capture of slow neutrons. The transmission efficiency I_{pB}/I_s of both sizes of instruments is low (it is 10^{-6} at 100 kev and 10^{-8} at 1 Mev for the 7·7 metre spectrometer). The sources studied with them are preferably 100 mc or greater.

The transmission efficiency of the 7·7 metre spectrometer is 4 to 6 times that of the 2 metre instrument, although both instruments have approximately the same solid angle Ω subtended by the crystal at the source, $\Omega = 3 \times 10^{-5}$. The bent crystal in the 7·7 metre spectrometer has greater diffraction efficiency P_H/ω because the diffraction line width ω is about 1·7 times smaller than for the 2 metre instrument. In addition the diffraction power P_H of the crystal in the larger spectrometer is three times greater because it is 6 mm rather than 2 mm thick.

The diffraction efficiency of line focus transmission spectrometers decreases with increasing γ-ray energy (at high energies P_H decreases roughly as E^{-2}, equation (2.9)). Also the γ-ray background increases with decreasing Bragg angle θ_B above 1 Mev (the background originates from γ-rays scattering from the Soller slit walls and is a function of the source strength and the intensity distribution in the spectrum). These two factors combine to limit measurements with this type of spectrometer to energies below about 1·3 Mev. In order to measure higher energies, it would be necessary to improve the Soller slit system and advantageous to use crystals which have greater reflecting power than quartz. The present instruments use slit systems with an angular resolution of 6 to 8 minutes of arc. A Soller slit with a 3 to 4 minute angular aperture could probably be constructed.

2.2.2. *The line focus reflection spectrometer.*

This crystal spectrometer is shown in figure 2.5. The γ-rays emitted from a line source S_1 or S_2 are Bragg diffracted from the bent crystal Q and refocused on the focusing circle at F_1 and F_2 respectively. The geometrical focusing conditions for this spectrometer have been considered in detail elsewhere (DuMond and Kirkpatrick 1930, DuMond 1958 a). The crystal diffraction power P_H for low energy γ-rays is greater for diffraction by reflection than by transmission. Also the radiation detector for the reflection spectrometer, if placed at the focus of the diffracted beam, can be small in comparison with the broad detectors required by the transmission spectrometers and so have a relatively low background counting rate. A disadvantage of the spectrometer if used for high energy γ-ray measurements where the diffraction angle is small (probably the main reason why it has not received further consideration) is that the line source and the focus of the

diffracted radiation are close to the crystal. It would be difficult with this spectrometer to shield the source from the detector under these conditions. The spectrometer has been used successfully for the measurement of low energy x-rays.

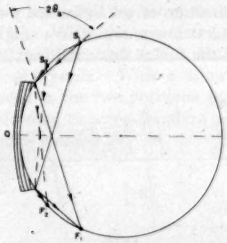


Figure 2.5. The reflection line source crystal spectrometer. The crystal Q is ground and bent to the focusing circle so that line sources S_1 and S_2 on the circle will refocus after diffraction at points F_1 and F_2 respectively. A γ-ray of energy E_γ diffracts from the crystal planes through an angle θ_B. The position of the line source and line image on the focusing circle are determined by the γ-ray energy and the d_H spacing of the crystal according to equations (2.1) and (2.2).

2.3. *Broad Source Transmission Spectrometers*

Two types of broad source spectrometer will be discussed, the flat crystal transmission and the focusing transmission (Cauchois) spectrometers. Both have a lower transmission efficiency than line source spectrometers, but may have comparable peak counting rates at the detector if extensive (strong) γ-ray sources are used. Such sources are obtained by neutron capture in a large sample placed in a thermal reactor, or by particle excitation in the relatively broad target of a high current charged particle accelerator.

2.3.1. *The flat crystal diffraction spectrometer.*

A flat crystal transmission spectrometer was constructed by Knowles (1959) to measure γ-rays resulting from neutron capture in a broad source in a thermal reactor. It has low geometrical efficiency, 10^{-9} to 10^{-11}, but excellent resolution, greater than 0·5%, and has been used to measure γ-rays from 65 kev to 5 Mev. For measurement over this range of energies the spectrometer is operated either as a single or as a double crystal device and for optimum efficiency uses different diffracting crystals to measure different parts of the energy spectrum. The proper choice of crystals depends on the intensity and detail of the spectrum under consideration.

As a double crystal device, γ-rays are diffracted on transmission through a single flat crystal, then further diffracted through a second crystal set in the antiparallel position and finally detected by a large NaI scintillator D, figure 2.6(*a*). The relative angle $2\theta_B$ between the reflecting planes of the two crystals is measured and the γ-ray energy calculated from equations (2.1) and (2.2).

The doubly diffracted γ-rays traverse three lead baffles (Soller slits) whose purpose is to shield the NaI detector from the direct radiation (at 1 Mev the Bragg

diffraction angle is 6 minutes of arc when diffracting from the (211) planes of calcite). The Soller slit $C_1 C_2 C_3$ has an angular divergence of 45 seconds. It is coupled to the crystal turntables through an optical system, figure 2.6 (*b*), employing a parallel light beam and photocells which keep it automatically oriented within a few seconds of arc of the direction of the diffracted γ-ray beam.

The spectrometer is used at low energies as a single crystal device employing one of the two crystals and the Soller slits. Diffracting from the (211) planes of

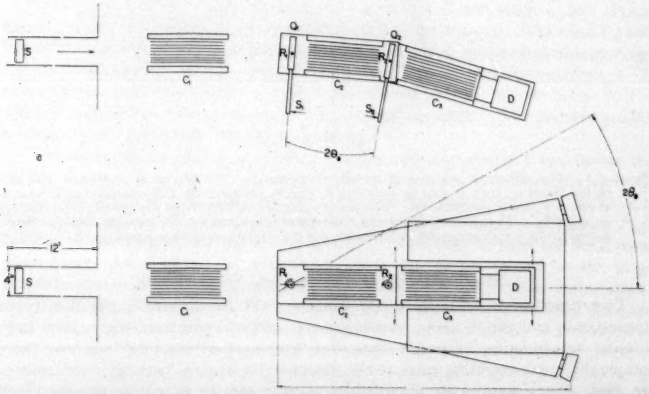


Figure 2.6 (*a*). Schematic drawings of the mechanical system of the flat crystal spectrometer. The spectrometer includes the two crystals Q_1 and Q_2 mounted on two precision turntables S_1 and S_2 which rotate about centres R_1 and R_2 respectively. The radiation from a broad source S proceeds through Soller slit C_1. Additional Soller slits C_2 and C_3 shield the direct radiation from the detector D, a NaI(Tl) scintillator.

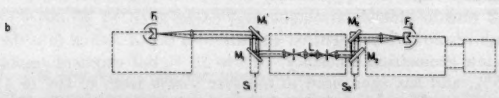


Figure 2.6 (*b*). Two optical lever systems are used to keep the slit system properly oriented with respect to the turntables S_1 and S_2. Parallel light beams originate from the common source L, reflect from the mirrors $M_1 M_1'$ and $M_2 M_2'$, fixed to the turntables, and are focused on optical locating devices F_1 and F_2 (F_1 and F_2 consist of a split prism and two photocells in a direct current bridge circuit).

calcite the spectrometer has resolution 0·6 to 6% between 100 kev and 1 Mev (the resolution is determined by the 45 second angular width of the Soller slit system) but relatively high transmission efficiency. For the measurement of γ-ray energies below 500 kev high order reflection planes such as the (633) and (844) planes of calcite are used to obtain improved precision with a factor of 3 and 4 times better resolution. The single flat crystal spectrometer is used frequently at poor resolution for preliminary scanning of an unfamiliar spectrum.

2.3.1.1. *Resolution and precision of energy measurements.*

For precision measurement of γ-ray energies the flat crystal spectrometer measures the same γ-rays diffracted on the two sides of the direct beam, positions A_L and A_R, figure 2.7 (*a*) (the same procedure that is used with line source spectrometers). The line shapes are measured by using the diffracting crystals in their parallel position, figure 2.7 (*b*). With this geometry the spectrometer has zero dispersion and the shape of the diffracted line is attributed only to the mosaic distribution of the diffracting crystals. With a knowledge of the line shape a complicated spectrum observed in the two positions A_R and A_L can be fitted with a precise energy scale by reference to some standard γ-ray and in addition it can be analysed into its component γ-rays.

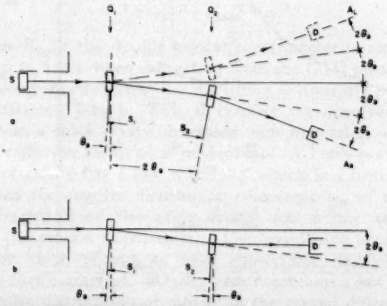


Figure 2.7 (*a*) and (*b*) illustrate the radiation path of the double crystal spectrometer with crystals arranged, (*a*) in the anti-parallel position, and (*b*) in the parallel position.

Figure 2.8 shows the titanium spectrum (Knowles 1959) between 140 kev and 5·0 Mev plotted on a linear wavelength scale in MÅ. The parts of the spectra A and A′ were measured by diffraction from the (211) planes of calcite with the double crystal arrangement and the part B was measured with a single calcite crystal. The widths of the double crystal diffracted lines at half-maximum intensity are about 0·4 to 0·6% of their energy between 0·34 and 5 Mev. Measurements of γ-ray energies of ^{49}Ti and ^{24}Mg relative to the annihilation radiation† are listed in table 2.1. The precision of these measurements is limited to 0·1% because of uncertainties in the linearity of the tangent screws which provide the rotation for the crystal turntables. Comparative γ-ray energy measurements in the region 1 to 2 Mev, to better than 0·01%, are expected when full use is made of the standard scale with optical system planned for this spectrometer (Knowles 1959).

† It has been found convenient to refer all energy measurements made with the flat crystal spectrometer to the annihilation γ-ray energy which is present as part of the background in all neutron capture γ-ray spectra. This energy has been measured by crystal diffraction relative to known x-ray energies. The ' best measured value ' of this annihilation energy is quoted to 1·3 parts in 10^4 (Muller *et al.* 1952), table 2.1.

The resolution R of the single flat crystal transmission spectrometer is defined by equation (2.4). The corresponding resolution R_D of the double flat crystal spectrometer is (in agreement with the more general definition given in the footnote †

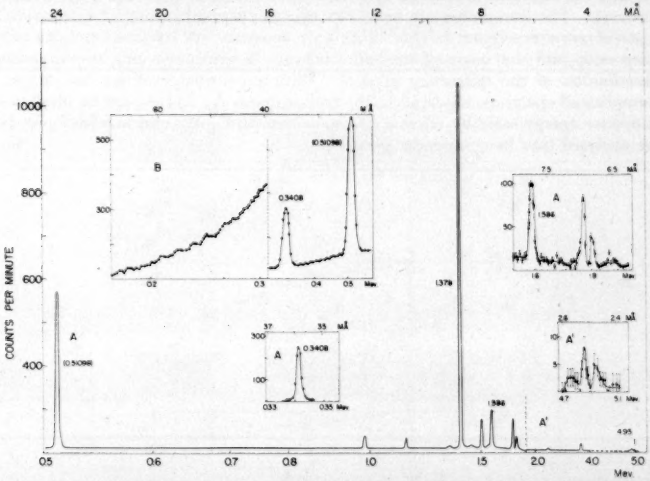


Figure 2.8. Part of the γ-ray spectrum of titanium (predominantly ^{49}Ti following neutron capture in ^{48}Ti) in the energy range 0·14 to 5·0 MeV. The abscissae are on a linear ångström scale. For the measurement of this spectrum, three spectrometer arrangements were used. The γ-rays in the region A, 0·5 to 1·80 MeV, were diffracted during transmission from the (211) planes of two 23 mm thick calcite crystals. For the region A′, 1·80 to 5·0 MeV, it was more efficient to use two calcite crystals each 62 mm thick. The thicker crystals are used with the best efficiency below 1·80 MeV because γ-ray absorption in the crystals is excessive and because broadening of the diffracted γ-ray line reduces the energy resolution. In region B, 0·14 to 0·52 MeV, the single crystal arrangement was used, reflecting the radiation from the (211) planes of a single calcite crystal. The diffracted radiation is mainly continuum on which is superimposed the 0·3408 MeV γ-ray together with a large number of weak γ-rays. All γ-ray energies are measured relative to the energy of the annihilation radiation, 0·51098 MeV.

on p. 462) expressed as the ratio of the observed double crystal diffraction line width ω_D at half maximum intensity and half the total scattering angle $2\theta_B$:

$$R_D = \Delta E_\gamma / E_\gamma \simeq \omega_D / 2\theta_B. \qquad \ldots\ldots (2.14)$$

The observed width ω of the diffracted line in the single flat crystal spectrometer is a function of the Soller slit angular aperture ω_B, the crystal diffraction width ω_M' and the radiation width ω_γ. The observed width may be given approximately in terms of the separate widths, equation (2.3). The single flat crystal spectrometer, described above, has a Soller slit aperture width $\omega_B \simeq 45$ seconds of arc. The flat crystals used with this spectrometer are nearly perfect calcite or quartz and for these the crystal diffraction width ω_M' is only a few seconds of arc. Since

the radiation width ω_γ is negligible for most γ-rays the observed diffraction line width $\omega = \omega_B$ and is therefore independent of the γ-ray energy. The double flat crystal spectrometer has a diffraction line width ω_D which is usually much narrower than the Soller slit width ω_B. This is possible with the two crystal spectrometer because the angular measurements are made by rotating the second crystal relative to the first, the first crystal remaining fixed relative to the axis of the Soller slit. Consideration of the ' widths ' illustrated in the diagrammatic spectrometer of figure 2.1 (*b*) will make it clear that only the diffraction widths ω_{M1}' and ω_{M2}' respectively of the two crystals and the radiation width ω_γ can contribute to the observed width ω_D even when other widths are large in comparison. For the double crystal spectrometer the observed width of a diffraction line is given by the approximate equation

$$\omega_D^2 \simeq \omega_{M1}'^2 + \omega_{M2}'^2 + \omega_\gamma^2. \qquad \ldots\ldots (2.15)$$

The resolution R_D of the double crystal spectrometer is almost independent of energy from 0·5 to 5 MeV when diffracting from the (211) planes of thick calcite crystal (with focusing spectrometers the resolution is inversely proportional to the energy of the diffracted γ-ray). This is because the observed width ω_D of a diffraction line from a thick crystal decreases with increasing γ-ray energy (it is assumed that the radiation width ω_γ is negligible). A 1 MeV γ-ray diffracted from the (211) planes of calcite has a line width ω_M' which is a factor of two to three times broader than the angular distribution of mosaic ω_M in the crystal, but a 3 MeV γ-ray diffracted from the same crystal has a line shape very nearly representative of the mosaic distribution. The broadening occurs for diffracting crystals which are thick enough to have appreciable multiple scattering and pronounced secondary extinction. Under such conditions a few crystallites in the wings of the angular distribution of mosaic in the crystal diffract the γ-ray with almost the same efficiency as the many crystallites at the centre of the distribution.

Since the resolution of the flat crystal spectrometer is inversely proportional to the diffraction angle θ_B for a constant diffraction line width ω_D, equation (2.14), it can be decreased by diffraction at larger angles, from higher order crystal planes. For example, the 1·378 MeV γ-ray of ^{49}Ti has been measured with a resolution to 0·08% by diffracting from the (844) planes of calcite in the double crystal arrangement. The spectrometer efficiency is greatly reduced when using high order crystal reflections so that great improvement in resolution is practical only for the more intense γ-rays.

2.3.1.2. *Efficiency and range of energy measurements.*

The transmission efficiency I_{TS}/I_s' of the single flat crystal transmission spectrometer is the same as that of the diagrammatic spectrometer already discussed. Its efficiency is described in terms of the crystal diffraction efficiency P_H'/ω and the angular width ω_B and ϕ_B by equation (2.11) (for this type of spectrometer ω_B is the angular aperture of a single slot of the Soller slit). To simplify the discussion equation (2.11) is rewritten in terms of the solid angle $\Omega_M = \phi_v \omega_M'/4\pi$ where ϕ_v is the vertical angle (normal to the plane of incidence) at the source subtended by the diffracting crystal and ω_M' is the crystal diffraction

width. The single crystal transmission efficiency I_{pS}/I_s' is rewritten from equation (2.11) in the form

$$\frac{I_{pS}}{I_s'} \simeq \Omega_M C_1 \frac{\omega_B}{\omega} \frac{P_H}{\omega_M'}. \qquad \ldots\ldots(2.16)$$

The fractional transmission of the Soller slits $C \simeq 0{\cdot}5$.

The transmission efficiency I_{pD}/I_s' of the double flat crystal spectrometer is less than that of the single crystal efficiency by a factor, the reflection efficiency P_H/ω_D of the second crystal relative to the first. The double crystal transmission spectrometer efficiency is given approximately by the expression

$$\frac{I_{pD}}{I_s'} \simeq \Omega_M C_1 \frac{\omega_B}{\omega} \frac{P_H}{\omega_M'} \frac{P_H}{\omega_D}. \qquad \ldots\ldots(2.17)$$

For both broad source spectrometers I_s' is the effective strength of the source (the effective surface area of the source facing the first diffracting crystal is equal to the face area of the crystal).

The transmission efficiency of both the single and double flat crystal spectrometers depends primarily on the solid angle Ω_M which is about 1 part in 10^{-8} when the diffraction width of the crystal ω_M' is two seconds of arc and the vertical divergence of the incident beam $\phi_v \simeq 1°$. Other factors affecting the transmission efficiency, such as the crystal diffraction efficiencies P_H/ω_M' and P_H/ω_D are large because the crystals used are nearly perfect and very thick, 1 to 7 cm. The crystal diffraction efficiency has a maximum value of $\frac{1}{2}$ (neglecting γ-ray absorption in the crystal). Close to the maximum conditions are reached for the diffraction by transmission of 1 Mev γ-rays from the (211) planes of 23 mm thick calcite crystals. Under these conditions the reflecting power P_H of the crystal is given by equation (2.7).

The transmission efficiency decreases from 10^{-9} to 10^{-11} between 1 and 3 Mev because the crystal reflecting power decreases with increasing γ-ray energy as shown in figure 2.2. The transmission efficiency of the single crystal spectrometer can be improved for the diffraction of low energy γ-rays with little loss in resolution (the resolution for the spectrometer is determined by the Soller slit angular aperture) by employing crystals which have a diffraction line width ω_M' (using crystals with a large distribution of mosaic crystallites) approximately equal to the Soller slit angular aperture ω_B.

While γ-ray energy measurements are made with the two crystals in the anti-parallel geometry, the double crystal spectrometer is used in the parallel arrangement for the absolute intensity calibration of the instrument. This is determined from measurements of the reflecting powers of the two crystals, Q_1 and Q_2. It is obtained for crystal Q_1 (or Q_2) from the ratio of two intensities, the intensity of a γ-ray after double diffraction from crystals Q_1 and Q_2 to the intensity of the same γ-ray after diffraction from Q_1 (or Q_2). Experimental values of the reflecting powers for the (211) planes of two calcite crystals are compared in figure 2.9 with the theoretical curves Q_1 and Q_2, assuming no primary extinction in the diffracting crystals. The theoretical curves have been calculated on the assumption that the single crystal diffraction line widths ω_M' above 3 Mev from 23 mm thick calcite crystals are the crystal mosaic distribution widths ω_M.

The Chalk River flat crystal spectrometer, as a single crystal instrument, is designed for the measurement of γ-ray energies as low as 20 kev but as yet has not been used below 65 kev. Strong self absorption of γ-rays in the source, absorption in the diffracting crystals, and various materials in the flight path have reduced the intensity of the low energy γ-rays at the detector and made their measurement much more difficult than that of 1 to 2 MeV γ-rays. At high energies, greater than 4 MeV, the diffraction efficiency of calcite decreases rapidly, figure 2.9, and in addition the direct radiation is scattered from the Soller slit walls into the γ-ray detector (a 5 MeV γ-ray, diffracted from the (211) planes of calcite, has a diffraction

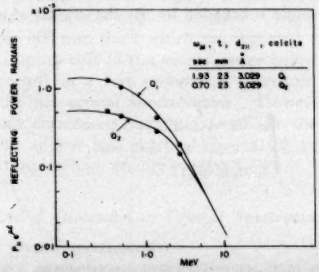


Figure 2.9. The reflecting power P_H divided by the factor $\epsilon^{-\mu d}$ (the fractional absorption of γ-rays in the crystal) plotted as a function of γ-ray energy for the (211) planes of two calcite crystals Q_1 and Q_2 each 23 mm thick. The lines through the points are the calculated reflecting powers assuming a Gaussian shape for the distribution of mosaic in the crystals. The widths of the mosaic distributions in the crystals Q_1 and Q_2 are $\omega_{M_1} = 1{\cdot}93$ and $\omega_{M_2} = 0{\cdot}70$ second of arc respectively.

angle of about 1 minute of arc). By constructing a Soller slit with higher resolution or employing crystals with more reflecting power so that higher order diffraction is practical, the range of the flat crystal spectrometer can probably be increased above 5 MeV.

2.3.2. *The Cauchois spectrometer.*

This is a focusing spectrometer which uses a broad source and focuses the γ-rays diffracted from the bent crystal Q, figure 2.10, to a line. For high energy measurements the source of radiation S is placed behind a low resolution Soller slit C, and γ-rays of different energies emitted by it are diffracted and focused at different points F_{E_1} and F_{E_2}, forming a focusing circle. The γ-ray spectrum is recorded at the focusing circle on a photographic plate.

The resolution of this spectrometer is equivalent to that of a line source spectrometer which uses a bent crystal of the same radius of curvature if the spread in the focused line in the photographic emulsion can be neglected. The bent quartz crystal used by Chupp *et al.* (1957 a, b), with a two-metre radius of curvature, has a diffraction line width of about 24 seconds of arc. Earlier workers (Cauchois 1932 and Frilley *et al.* 1951) have used this type of spectrometer, employing bent mica crystals, with radii of curvatures of about 29 cm and γ-ray

line widths of 15 seconds of arc at the focus. These have been used principally for the study of low energy x-rays emitted by radioactive elements. Since the photographic method permits large parts of the spectrum to be detected simultaneously systematic errors inherent in the point by point method of measurement should be absent. For precision measurement the film is calibrated with known γ-rays and x-rays. In table 2.1 are listed a few of the best measurements made with this spectrometer (Chupp *et al.* 1957 a, b).

The peak counting rate for a γ-ray diffracted by the Cauchois spectrometer is comparable with that of the single flat crystal spectrometer if the diffracting crystals are identical and if the solid angle subtended by the bent crystal at its focal point equals the solid angle subtended by the flat crystal at a point on the source.

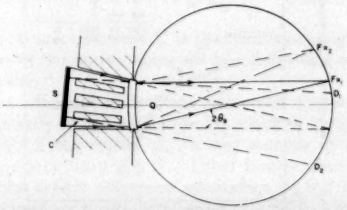


Figure 2.10. The Cauchois broad source transmission spectrometer. A γ-ray from the source S with energy E_γ diffracted at an angle θ_B to the planes of a bent crystal focuses at F_{E_1} on the focusing circle. A γ-ray of different energy will focus at a different point, F_{E_2}, on the circle. A baffle system C is used to confine the direct radiation between D_1 and D_2.

The factors determining the peak counting rate of a diffracted γ-ray are those of the single crystal spectrometer and are given in equation (2.16). At low γ-ray energies, less than 300 kev, the transmission efficiency of the Cauchois spectrometer equals that of the single flat crystal spectrometer and is about 1 part in 10^{-9}. As expressed by equation (2.16) the efficiency is independent of the crystal diffraction width ω_M'.† The resolution R of the spectrometer is the same as for the line focus spectrometer, equation (2.4), and the observed diffraction line width ω at half-maximum intensity is a function of similar separate widths: the crystal diffraction width ω_M', the aberration width ω_f and the radiation detector width ω_s (which replaces the source width). Since the line width ω may be reduced by using thick crystals bent to a large radius of curvature (as was the case for the 7·7 metre Argonne spectrometer) it would appear advantageous to construct a large Cauchois spectrometer. However, since the focusing takes place in a plane the intensity of the diffracted γ-ray at the photographic film or plate is proportional to R^{-1} where R is the radius of curvature of the bent crystal. The upper limit of size for the spectrometer is determined by the strength of the available source. For the variation of the mosaic distribution width with the radius of curvature of the bent crystal refer to the footnote † on p. 462.

† For this spectrometer the defining Soller slits are used only for shielding purposes. In applying equation (2.16) to this spectrometer the Soller slit width ω_B equals the line width ω. The diffracted intensity is the same whether it comes to a focus or not. The focusing is required only to separate spatially radiations of different energy.

At energies above 300 kev this spectrometer has only moderate transmission efficiency (similar to other transmission focusing spectrometers) because it uses thin crystals of quartz and mica. A particular disadvantage of this spectrometer is that the radiation scattered out of the direct beam by the slit system at small diffraction angles is greater than for a line focus or flat crystal spectrometer. However for γ-ray measurement below 350 kev the Cauchois spectrometer has definite advantages. The γ-rays are focused to a point prior to detection so that the ratio of γ-ray peak intensity to extraneous background is high (despite the lack of energy discrimination in the photographic plate).† A further advantage of the Cauchois spectrometer, mentioned previously, is the simultaneous recording of large portions of the spectrum at a single exposure. Chupp *et al.* (1957 a, b) used nuclear emulsion plates 600 mm thick which are reported to be 20% efficient for 80 kev γ-rays. This implies that in the energy range below 300 kev the Cauchois spectrometer could record in a few minutes a spectrum which would require several hours using the flat crystal spectrometer. However, the film efficiency, crystal reflectivity, and extraneous radiation from the direct beam combine to make this spectrometer appear less practical than a flat crystal spectrometer for the measurement of complex high energy spectra.‡

2.4. *A Discussion of Crystal Spectrometers*

2.4.1. *Comparison of crystal spectrometers.*

The usefulness of a spectrometer for γ-ray measurements is determined not only by definite characteristics, its efficiency and resolution, but is also determined by less definite factors such as source strength, extraneous background, and instrument stability which may change with the experiment. For this reason no general comparison is possible. However, a particular comparison will be made of three instruments, the 7·7 metre bent crystal (line source) spectrometer and the single and double flat crystal (broad source) spectrometers used for the study of neutron capture γ-rays.

These spectrometers operate under similar conditions; both line and broad source are produced in a similar flux of slow neutrons, the solid angle subtended at the source by the crystal is the same, about 3×10^{-5}, and their source-to-crystal distance is approximately 8·0 metres.

In comparing their peak counting rates I_p it is assumed, in addition to the conditions mentioned above, that the height of the line source and the height and breadth of the broad source subtend angles of 1° with their respective crystals, and that the breadth of the line source subtends an angle of 10 seconds at the bent crystal (an angle which is somewhat less than the minimum observed diffraction line width of 13 seconds). Under these conditions the ratio of the strengths of line and broad source is about 1 to 100. The transmission efficiency of the 7·7 metre bent crystal spectrometer is much greater than that of the single flat crystal spectrometer.

† The extraneous radiation occurs over a range of energies and with NaI scintillation detectors a large fraction of it can be eliminated by suitable electronic discrimination.

‡ Recently Chupp *et al.* (1958), using a 20 curie source of ^{56}Co and suitable baffles to shield the photographic plate from the detector, have observed the 1·173 and 1·332 MeV γ-rays and more recently (1959), using a 1500 curie source following the reaction ^{1}H(n, γ) ^{2}H, they have measured the 2·224 MeV γ-ray of deuterium with the 2 metre Cauchois spectrometer.

In particular, if the flat and bent diffracting crystals had the same reflecting power, the ratio of their efficiencies would be about 100 to 1 so that the spectrometers would have equal counting rates. In practice, the flat crystal spectrometer has the greater counting rate for γ-ray energies above 0·5 MeV because it uses a 23 mm thick calcite crystal rather than a 6 mm quartz crystal. This is shown in figure 2.11 where

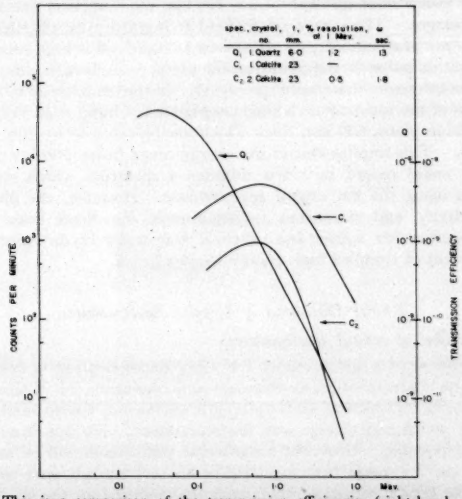


Figure 2.11. This is a comparison of the transmission efficiencies (right-hand scales) and peak intensity counting rates (left-hand scale) as a function of γ-ray energy of diffraction spectrometers used to study γ-rays following slow neutron capture in sources placed in thermal neutron reactors. The ordinate Q refers to line source spectrometers and C to flat crystal spectrometers. The 7·7 metre bent crystal spectrometers Q_1, and the single and double flat crystal spectrometers C_1 and C_2 respectively are compared. For all spectrometers the solid angle at the source subtended by the first crystal is chosen as 3×10^{-6} (which represents closely the solid angles of the actual instruments). The peak counting rates are representative of the flat crystal spectrometers. The counting rate for the single broad source spectrometer is equal to that of the line source spectrometer when equivalent diffracting crystals are used.

the curves Q_1 and C_1 are the calculated peak counting rates, left-hand scale, and transmission efficiencies, right-hand scale, as a function of γ-ray energy for the 7·7 metre bent crystal and single flat crystal spectrometers respectively. The peak counting rate of the double flat crystal spectrometer, curve C_2, is less than that of the single flat crystal spectrometer but better than that of the bent crystal spectrometer in the region of 1 to 3 MeV. At low energies the peak counting rate for each spectrometer is very nearly constant (the slight decrease with decreasing energy is caused by the increased absorption of the incident radiation in the diffracting crystals). At these energies the counting rate for the bent crystal spectrometer is greater than that of the single flat crystal spectrometer by a factor of ten because the mosaic distribution of the bent crystals is proportionately greater.

The above and other comparisons of operating spectrometers are of limited value because very considerable improvements are expected from the use of more efficient crystals. These possibilities will now be discussed.

2.4.2. *The improvement of spectrometer efficiency.*

At low γ-ray energies an improvement in transmission efficiency of the existing bent crystal spectrometers may be made, equation (2.13), by increasing the solid angle Ω subtended by the crystal at the source. This may be done (still keeping the focusing aberrations small) either by increasing the crystal size or by increasing the curvature of the bent crystal. Large crystals (quartz) are available but are expensive and their use implies the construction of comparably large Soller slits

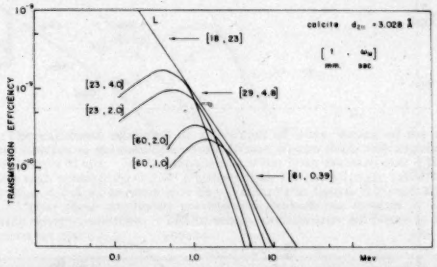


Figure 2.12. The calculated efficiency (as a function of γ-ray energy) for the double flat crystal spectrometer employing calcite crystals 23 and 60 mm thick with mosaic distributions of 1·0, 2·0 and 4·0 seconds of arc, and diffracting from the $d_{211} = 3{\cdot}028$ Å planes. The maximum permissible spectrometer efficiency as a function of γ-ray energy is given by curve L. The crystal thickness and mosaic distribution width for maximum efficiency at a particular γ-ray energy is indicated at a number of points. In all calculations of efficiency it is assumed that the first crystal subtends a solid angle $\Omega = 3 \times 10^{-6}$ at the source.

and detector. The use of crystals bent to a small radius of curvature increases the transmission efficiency but may decrease the counting rate because sources of smaller geometrical size must be used with the smaller instruments.

For γ-ray energies greater than 100 kev the spectrometer efficiency can be increased by improving the reflecting power of the diffracting crystals. For the flat crystal spectrometer this may be done by increasing the crystal thickness (see equations (2.16), (2.17) and (2.9)). This improvement is limited by the γ-ray absorption in the crystal, and the limit is reached when the crystal thickness t equals μ^{-1} for single-crystal and $0{\cdot}5\mu^{-1}$ for two-crystal spectrometers where μ is the linear absorption coefficient for the crystal. A crystal of thickness μ^{-1} will have a reflecting power P_H, for high energy γ-rays diffracted by transmission, given by (see equation (2.9)),

$$P_H \simeq r_0^2 \lambda^3 d_H (F_H/v)^2 \mu^{-1} e^{-1}. \qquad \ldots\ldots (2.18)$$

The improvement in efficiency at high energies, obtained by increasing the thickness of the calcite crystal, is illustrated in figure 2.12. The curves have been

calculated for a double crystal spectrometer using crystals with angular distributions of mosaic ω_M of 1, 2 and 4 seconds of arc and thicknesses of 23 and 60 mm. It is profitable to use different crystal thicknesses for different energy ranges of a spectrum as was done for the titanium spectrum, figure 2.8.

The transmission efficiency of a crystal spectrometer can also be improved by replacing the quartz or calcite by crystals of higher electron density. The factor $(F_H/v)^2\,\mu^{-1}$ in equation (2.18) is sensitive to such changes. The calculated points shown in figure 2.13 are for a number of crystals of different atomic number Z for

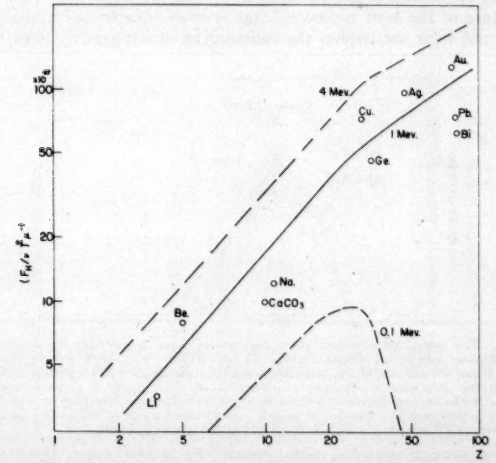


Figure 2.13. The expression $(F_H/v)^2\,\mu^{-1}$ plotted as a function of the atomic number Z for a number of crystals to illustrate the increase of crystal reflecting power with Z. The points have been calculated for a 1 Mev γ-ray and are represented by a smooth line. The broken lines show the mean reflecting powers at 4 and 0·1 Mev. In the calculation of the structure factor F_H the interplanar spacing of the crystal d_H is taken arbitrarily as 2·00 Å and the radiation scattered from the atoms of the crystal unit cell are in phase.

a γ-ray of 1 Mev. A representative curve is drawn through these points. Representative curves for 4 and 0·1 Mev γ-rays are also shown. Calcite has relatively poor reflecting power compared with some crystals of higher Z in the range of γ-ray energies 1 to 4 Mev. This is because the crystal coherent scattering increases with atomic number, roughly as Z^2, but the absorption of radiation only as Z (absorption takes place primarily by the Compton process in this energy range). At low energies, below 100 kev, the absorption increases roughly as Z^4 since fluorescent absorption is the dominant process. At low energies, crystals of high Z are less efficient than those of low Z, as shown by the lowest curve of figure 2.13. Artificially grown germanium crystals are only slightly less perfect than calcite. The use of such crystals, 23 mm thick, in the double crystal diffraction spectrometer

would improve its transmission efficiency by a factor of 30 in the energy range above 5 MeV, over its present efficiency (using 23 mm thick calcite crystals). A spectrometer employing two flat gold crystals, each of 2 second mosaic, would be even more efficient. Such crystals would have roughly constant reflecting power up to 5 MeV. However because of their crystal properties gold crystals may be

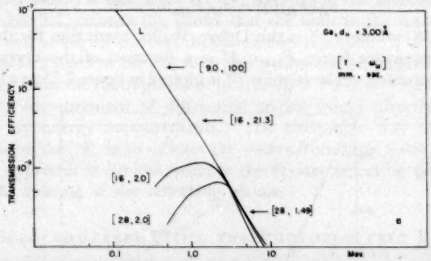


Figure 2.14 (*a*). The calculated efficiency as a function of γ-ray energy of the double flat crystal spectrometer which uses germanium crystals 16 and 28 mm thick with angular mosaic distributions of 2·0 seconds of arc. In the calculations it has been assumed that diffraction takes place from a plane with spacing $d_H = 3{\cdot}00$ Å (rather than for the strongly reflecting (220) planes of germanium, $d_{220} = 2{\cdot}00$ Å) in order that the efficiencies in figures 2.14 and 2.12 may be directly compared. Under these conditions spectrometers which use crystals of comparable mosaic have the same energy resolution. The optimum spectrometer efficiency as a function of γ-ray energy is given by curve L.

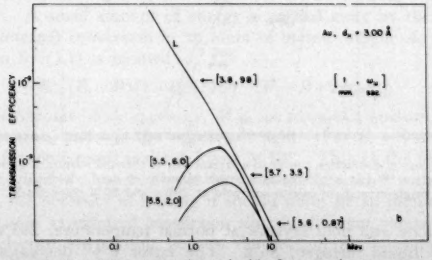


Figure 2.14 (*b*). The corresponding curves for a double flat crystal spectrometer using gold single crystals 5·5 mm thick and mosaic distributions of 2 and 6 seconds of arc.

difficult to grow with a small mosaic distribution. In figure 2.14 (*a*) and (*b*) are shown double flat crystal spectrometer transmission efficiencies as a function of γ-ray energy when diffraction takes place from germanium and gold single crystals. For these calculations the spacings of the diffracting planes d_H are taken to be 3·00 Å in order that direct comparison can be made with the spectrometer using calcite crystals, figure 2.12.

Although near perfect (1 to 2 second mosaic distribution) crystals of very high electron density are not obtainable at the present time, it may be possible to produce

high density crystals with mosaic distribution of the order of 45 seconds of arc. For the diffraction of γ-rays below 5 MeV, these crystals would have ample reflecting power and this could be sacrificed in order to improve the resolution by diffraction from higher order planes. The change of reflecting power with decreasing crystal plane spacing is obtained from the factor

$$d_H(F_H/v)^2 = d_H(F_{H0}/v)^2\,e^{-4M} \qquad \ldots\ldots(2.19)$$

of equation (2.18) where e^{-4M} is the Debye–Waller correction for thermal motion to the crystal structure factor F_{H0}. M is a function of the crystal spacing d_H and of the temperature. The factor e^{-4M} is plotted in figure 2.15 as a linear function

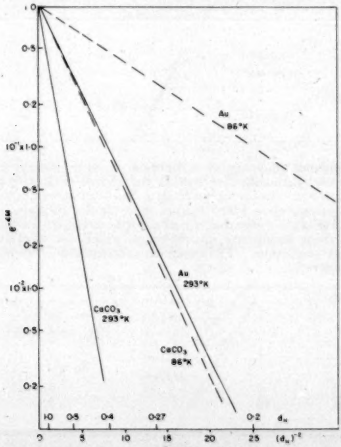


Figure 2.15. The loss in diffraction power of a crystal due to thermal motion is given by the Debye–Waller factor, e^{-4M}. It is plotted as a linear function of d_H^{-2} where d_H is the crystal plane spacing for single crystals of calcite and gold at 293°K and 86°K respectively.

of d_H^{-2} for calcite and gold crystals at normal temperature, 293°K, and at the temperature of liquid nitrogen, 86°K. The factor e^{-4M} decreases slowly with decreasing d_H for d_H greater than 1 Å but decreases by a factor of about 10^{-3} for a crystal at room temperature between $d_H = 1{\cdot}0$ Å and $d_H = 0{\cdot}25$ Å. However, much of the coherent scattering lost through thermal motion in the crystal may be recovered by cooling the crystal to the temperature of liquid nitrogen.

2.4.3. *The precision of energy measurements.*

The fractional precision $\Delta E_\gamma/E_\gamma$ attributed to a γ-ray energy measurement E_γ (E_γ greater than 30 kev) made with a diffraction spectrometer is approximately equal to the fractional precision $\Delta\theta/\theta$ with which the scattering angle θ can be measured.

$$\Delta E_\gamma/E_\gamma \simeq \Delta\theta/\theta. \qquad \ldots\ldots(2.20)$$

In crystal spectrometry the measure of θ is made between the right- and left-hand diffracted line of an isolated γ-ray. No prior knowledge of the γ-ray line shape is necessary unless the γ-ray has a natural width which is comparable to the line width. For this reason the uncertainty $\Delta\theta$ in the position of the line (uncertainties in counting statistics are excluded since for reasonably intense lines the statistical error is negligibly small) is due solely to uncertainties in the mechanical measurement of θ. Table 2.1, column (8), shows that the limit in the measurement of θ is 0·2 to 0·3 second of arc for the best measurements reported. For 100 kev γ-rays this corresponds to 3 parts in 10^5 and for 1 Mev γ-rays can be expected to be 3 parts in 10^4 diffracting from the (310) planes of quartz $d_H = 1{\cdot}178$ Å. Any improvement in the precision measurement of diffraction angles would directly improve the precision of γ-ray energy measurements. The alternative way to improve the fractional precision $\Delta E_\gamma/E_\gamma$ is to diffract the γ-rays from high order crystal planes. However, this approach is limited because the crystal reflecting power decreases rapidly with the spacing of the reflecting planes.

§ 3. Spectrometers using the Photoelectric Effect

3.1. *Introduction*

In discussing spectrometers based on the photoelectric effect we shall consider internal conversion as well as the photoelectric effect proper. In either case the energy difference between the initial and final nuclear states E_t is very nearly equal to the sum of the kinetic energy of the emitted electron and the binding energy of the atomic shell from which the electron was removed. Most of these binding energies are known, from the study of x-ray absorption, with an accuracy to about 10 ev. A small amount of energy is carried away by the recoiling ion. Thus for the internal conversion in an atom of atomic weight A, an electron of energy E given by (3.1) is emitted:

$$E \simeq (E_t - B)[1 - \tfrac{1}{2}(AMc^2)^{-1}(E_t - B + 2m_0c^2)]. \qquad \ldots\ldots(3.1)$$

Here B is the electronic binding energy, M is the mass of a nucleon, m_0 is the rest mass of the electron, and c is the velocity of light. For $E_t = 4$ Mev and $A = 20$ the correction for ionic recoil is about 1 part in 10^4. Though the recoil correction is not always negligible compared with the precision attainable with some modern spectrometers, the necessity of making it should never be an important source of error. In the case of external conversion slightly different values of the energy of the ejected electron are possible, because the electron may be emitted at different angles to the incident photon direction. If the photon energy is sufficiently greater than m_0c^2, the maximum spread of electron energies is approximately $2(AMc^2)^{-1}E_\gamma^2$, E_γ being the energy of the photon. Thus the relative spread in electron energies in external conversion is of the same order of magnitude as the correction which must be applied in internal conversion. If the photon has been emitted by an atom of atomic weight A', the transition energy E_t is related to E_γ by means of

$$E_t = E_\gamma[1 + \tfrac{1}{2}(A'Mc^2)^{-1}E_\gamma]. \qquad \ldots\ldots(3.2)$$

Of the spectrometers which measure γ-energies by first converting to charged particles, those based upon the photoelectric effect are simplest, since it is necessary to measure only the energy of just one electron per photon. The greatest precision

in the measurement of electron energies in the range of tens of kev to a few Mev has usually been achieved with β-spectrometers using magnetic fields. If internal conversion is to be used, the source material is placed at the source position of the spectrometer; if external conversion is to be employed, a piece of material of high atomic number is fixed in the source position to act as a converter. A thorough review of magnetic β-spectrometers has been given by Siegbahn (1955). We shall be content to discuss in §§ 3.2 and 3.3 the principles of focusing of each of the two basic classes of devices, and to describe one modern representative of each class. In the final subsection, § 3.5, of this section we shall give some comparisons of the two types of spectrometer.

The counting rate C_1 in a spectrometer using internal conversion is proportional to the source strength S_1, to the internal conversion coefficient for the particular nuclear transition and atomic shell in question α, to the spectrometer transmission T, and to the efficiency of the detector ϵ_d; S_1 is proportional to A the area of the source, to ρ the density of the source material, to f the number of photons emitted per second per gramme of source material, and to t_1 the mean source thickness. Thus

$$C_1 = A\rho f t_1 \alpha T \epsilon_d. \qquad \ldots\ldots(3.3a)$$

There will be a similar formula for external conversion, but it will depend upon the geometrical arrangement of the source and converter. To facilitate discussion let us assume that the same β-spectrometer is used, that the source is replaced by a converter of thickness t_2 and of density ρ_2 of the same size, and that a cylinder of source material of length t and of the same cross-sectional area is placed behind the converter. Then the counting rate for external conversion C_2 is given by

$$C_2 = A\rho f t \bar{\Omega} \tau \rho_2 t_2 T \epsilon_d; \qquad \ldots\ldots(3.3b)$$

τ is the term in the mass absorption coefficient for the converter material corresponding to photoeffect in a particular atomic shell. $\bar{\Omega}$ is the solid angle subtended by the converter at a point in the source, divided by 4π and averaged over the whole source volume; in this crude calculation we ignore the angular dependence of the differential cross section for the photoelectric effect.

If a β-spectrometer is operated with fixed slits and baffles, the transmission T is independent of the electron energy; it is possible to make the detector efficiency ϵ_d very nearly constant for all electron energies of importance. Both counting rates, (3.3*a*) and (3.3*b*), vary strongly with the energy, however, because both α and τ decrease rapidly as the photon energy is increased. Thus a spectrometer based on the Compton effect or upon pair production should be expected to be superior at higher energies; some numerical comparisons are given in § 6. Note also that both α and τ increase very quickly as the Z of the atom concerned is increased.

At higher energies the resolution of a β-spectrometer depends mainly upon geometrical aberrations, which do not vary with the energy. At lower energies the energy lost by the electron in emerging from the source or converter becomes of increasing importance; as the photon energy is decreased below 1 Mev this effect increases very rapidly. The ultimate lower limit to the energy range in which β-spectrometers are useful at high precision is fixed by the difficulty in

preparing sufficiently thin sources. Very approximately, the most probable energy loss at a fixed initial energy is proportional to the superficial density of matter in the source or converter. It is thus exceedingly important, in work of high precision, to make very thin sources or converters which are also very uniform in thickness. In some spectrometers the necessary smallness of other source dimensions is also a cause of difficulty.

When the source material is available for insertion in the source position of the β-spectrometer one has the choice of which form of conversion to use. If the atomic number of the decaying nucleus is very low, and if a sufficiently large amount of the source material is available, one can sometimes get a greatly enhanced effect by external conversion in a converter of high atomic number. To get a rough numerical comparison of the two possibilities we shall use the arrangements described in deriving (3.3*a*) and (3.3*b*), with the added assumption that the sources and converter be circular in cross section and have diameter D. Let x stand for t/D. Then, for a right circular cylinder,

$$\overline{\Omega} = \tfrac{1}{2} + 2(3\pi x)^{-1}\{1-(1+x^2)^{1/2}[(1-x^2)E + x^2 K]\}. \qquad \ldots\ldots(3.4)$$

K and E are complete elliptic integrals of squared modulus $k^2 = (1+x^2)^{-1}$. The rate of increase of the product $t\overline{\Omega}$ for large values of t is sufficiently slow that not much is gained by making t greater than D. Thus for $t = D$ we find $t\overline{\Omega} = (0{\cdot}156\ldots)D$, and for t infinite we get $t\overline{\Omega} = (0{\cdot}212\ldots)D$; if the absorption of γ-rays in the source material were included, the contributions from large distances would be reduced still further.

If the formulae (3.3*a*) and (3.3*b*) are to be compared under conditions of comparable resolution, the superficial densities of the source in internal conversion and of the converter in external conversion must be roughly equal; thus we put $\rho t_1 = \rho_2 t_2$. With the assumptions outlined and with $t = D$ we obtain

$$C_2/C_1 = (0{\cdot}156\rho D\tau)/\alpha. \qquad \ldots\ldots(3.5)$$

Note that the presence of the factor D in (3.5) means that external conversion is more likely to be more efficient for larger spectrometers or for poorer resolutions; external conversion is also favoured by denser source materials, which allow more active material to be placed closer to the converter. In figure 3.1 (*b*) we show the variation with Z of the internal K-conversion coefficients for E1, E2 and M1 transitions for a photon energy of 0·3 m_0c^2; the numbers are taken from the work of Rose, see Appendix IV of Siegbahn (1955). The horizontal line on the same drawing is the value of $0{\cdot}156\rho D\tau$ for typical values of the parameters; we have used $\rho = 2$ g cm^{-3}, $D = 0{\cdot}2$ cm, and the value of τ appropriate to conversion in the K-shell of uranium (adapted from Appendix I of Siegbahn 1955). Note that at very low values of Z the advantage in using external conversion is considerable. In figure 3.1 (*a*) similar curves for a photon energy of 5 m_0c^2 are shown; the advantage of external conversion is here on the whole slightly greater. When absorption of the γ-rays in the source is taken into account, the straight lines of figure 3.1 are replaced by lines which curve toward the abscissa at higher values of Z. Although the departure from the straight line is not easy to calculate exactly, it is easily shown that the point at $Z = 82$ for a photon energy of 0·153 MeV

is certainly not reduced by more than a factor of 3 below the no-absorption value; at 2·56 MeV the departure from the straight line is completely negligible.

Though we believe that as far as intensity of the desired radiation is concerned external conversion is often very profitable, we do not intend to suggest that this is the only point to be considered. The background of photoelectrons and Compton electrons produced in the large source volume in the external conversion case can be exceedingly troublesome; there may also be nuclear β-particles. It is often necessary to insert an absorber of some relatively dense material of low atomic number (e.g. copper) between the source and the converter; the displacement of the source by the absorber lowers the value of $\bar{\Omega}$. The necessity of using

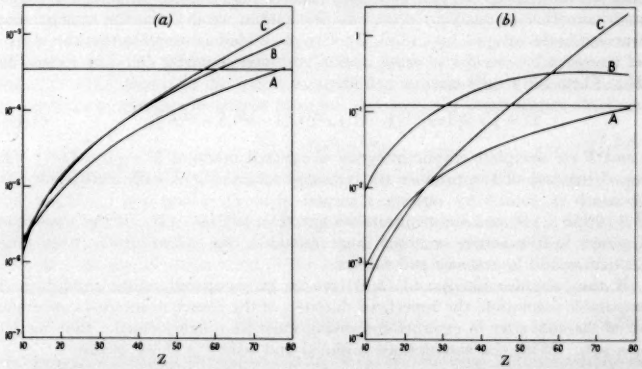


Figure 3.1. In A, B, and C we show the variations with atomic number of the internal conversion coefficients for electric dipole, electric quadrupole, and magnetic dipole radiations; (*a*) corresponds to a photon energy of 2·56 MeV, (*b*) to 0·153 MeV. In each case the straight line gives the value of $0{\cdot}156D\rho\tau$ for $D = 0{\cdot}2$ cm, $\rho = 2$ g cm^{-3}, and τ equal to the mass absorption coefficient for K-electrons in uranium.

very much stronger sources in the external conversion case results in considerable experimental difficulty; in particular, more shielding of the electron detector from direct γ-rays is required. Another point, which may be of importance in work of high precision, is that the intrinsic line-width is greater with external conversion. In addition to the uncertainty in the electron energy associated with ionic recoil, there is the fact that the natural width of the K-conversion line increases rapidly with atomic number; for uranium this natural width is of the order of 100 ev.

3.2. *Lens Spectrometers*

3.2.1. *Principles of focusing.*

The name 'lens spectrometers' is applied to that class of instruments in which the motion of the electrons is predominantly in the direction of the magnetic field. The solid angle in which electrons are accepted is symmetrically disposed with respect to an axis of cylindrical symmetry of the magnetic lines of force.

We shall begin our discussion by examining the motion of monoenergetic electrons emitted with momentum p from a point source situated in a uniform magnetic field of intensity H. The electrons which are ejected parallel to the field experience no magnetic force, and hence travel in a straight line ; those which are ejected perpendicular to the field move in circular orbits of diameter $D = 2p(He)^{-1}$, e being the electronic charge. Next consider an electron emitted at an angle θ to the magnetic field. The path of this particle is a helix of diameter $D \sin\theta$. The projection on a plane perpendicular to the field of a helical orbit is a circle. Let Θ be the angle subtended at the centre of this circle by the arc described in the first t seconds, let Z be the distance travelled in the direction of the field in the same time, and let r be the distance from the line of force passing through the source point after t seconds. Note that the magnitude of the velocity of the electron v is a constant of the motion. The helical orbits are described by the equations

$$r = D \sin\theta \left| \sin \tfrac{1}{2}\Theta \right| \qquad \ldots\ldots(3.6a)$$

$$Z = vt \cos\theta. \qquad \ldots\ldots(3.6b)$$

We shall use the symbol ψ for $\frac{1}{2}\Theta$. There is a useful geometrical interpretation of ψ ; it is the angle between the projections, on the plane normal to the field, of the initial velocity vector and the vector joining the source to the position of the electron at time t.

The projection of the electron on the normal plane moves about its circular orbit with a constant angular velocity, $\omega = v \sin\theta(\frac{1}{2}D \sin\theta)^{-1}$; thus

$$\psi = \tfrac{1}{2}\omega t = vD^{-1}t.$$

Eliminating the parameter t from the equations of motion, we find the following orbit equations :

$$r = D \sin\theta \left| \sin\psi \right| \qquad \ldots\ldots(3.7a)$$

$$\psi = Z(D \cos\theta)^{-1}. \qquad \ldots\ldots(3.7b)$$

Note that the first return of r to zero occurs after the particle has travelled a distance $\pi D \cos\theta$, equal to the pitch of the helix, in the Z direction. The orbits of electrons emitted with the same value of θ but at different azimuthal angles with respect to the line of force through the source lie on a surface of revolution. The curve formed by the intersection of this surface and a plane containing the axis of revolution is sinusoidal ; we shall use the term ' trace ' for this curve. The equation of the trace, $r = r(Z)$, may be found by eliminating ψ from (3.7 *a*) and (3.7 *b*).

We are now in a position to discuss the focusing property of a uniform magnetic field. Let us calculate the value of r at a distance $Z = L \equiv \pi D$ for a paraxial bundle of orbits. With the approximations appropriate to small values of θ one finds

$$r \simeq \tfrac{1}{2}L\theta^2. \qquad \ldots\ldots(3.8)$$

Because the distance r is of degree higher than the first in the small quantity θ, we have a focus at $Z = L$. Just as in the simple optical lens the aberration is third-order in θ ; expression (3·8) is a formula for the ' spherical aberration '.

Because the focal distance of the magnetic lens depends upon the electron momentum (it is proportional to p), the lens can be used as a spectrometer.

Assume that the field strength is adjusted so that the focal distance L for the electron group to be studied is equal to the distance between the source and a detector. Since the motion of particles for which $\theta = 0$ is completely unaffected by the field, it is necessary to insert an absorber to remove all electrons with trajectories for which θ is less than some minimum value, say θ'; this absorber can also be used to exclude radiations which could otherwise pass directly from source to detector. Because the aberration worsens the resolution if θ becomes too large, one should also use a baffle which rejects all electrons for which θ is greater than some maximum, say θ''. For electrons of the momentum in question these arrangements should produce a somewhat blurred image of the point source on a small circular aperture placed in front of the detector. If the momentum is larger or smaller than the nominal value, the image will be beyond or before the detector aperture; in either case the density of particles at the aperture is much reduced. If the strength of the magnetic field is varied through the nominal value for the electron group in question, the counting-rate will first rise and then fall, producing a 'line'. The field strength corresponding to the peak of the line depends upon the electron energy; thus the instrument is a spectrometer. The width of the line depends upon the limiting angles θ' and θ'', upon the size of the exit aperture, and upon the diameter of the source, since sources cannot be made negligibly small. If the source thickness is great enough, the energy loss suffered by the electron in leaving the source will contribute to the line width. In general both the line width and the counting-rate increase when any one of the four quantities we have mentioned is increased.

To make a spectrometer of the sort that we have been discussing it is not necessary to use a uniform magnetic field. Any system of lines of force possessing a cylindrical axis of symmetry has the ability to form an image, real or virtual; for further discussion and references to original papers consult the book by Zworykin *et al.* (1945). Various arrangements of coils have been used to construct lens spectrometers of divers types, see Siegbahn (1955). The calculation of the trajectories is more complicated with a non-uniform field, but it is not necessarily less precise. For the uniform field one can readily see that the image of a finite source is rotated through 360° and that the magnification is unity. These statements do not necessarily apply to lenses with non-uniform fields; spectrometers have been built to give different magnifications and image rotations.

One of the means whereby the efficiency of lens spectrometers has been improved is the adoption of the procedure called 'ring focus'. Though this type of focus has often been used with non-uniform fields, we shall again confine the discussion to the case of uniform fields. By examining formulae (3.7) one can see that electrons leaving the source at steeper angles to the field return to the spectrometer axis at shorter distances. Thus the traces corresponding to neighbouring values of θ must intersect. We are thus led to look for a pair of values of r and Z, r_1 and Z_1 say, and the associated angle θ_1, such that traces with values of θ in the neighbourhood of θ_1 are bunched together as they pass the point (r_1, Z_1); the restriction to small values of θ is now relaxed. In the neighbourhood of r_1 and for a fixed value of Z we have

$$r \simeq r_1 + \frac{\partial r}{\partial \theta}(\theta - \theta_1) + \dots; \qquad \dots\dots(3.9)$$

the partial derivative is to be evaluated for fixed Z, at $Z = Z_1$, $\theta = \theta_1$. The mathematical condition for a first-order ring focus is $\partial r/\partial\theta = 0$. Starting from (3.7) one obtains after a little manipulation

$$\psi_1^{-1} \tan \psi_1 = -\tan^2 \theta_1 ; \qquad \ldots\ldots(3.10)$$

ψ_1 is the value of the angle ψ, defined above, for $r = r_1$ and $z = z_1$. A table of the minimum ψ_1 satisfying (3.10) as a function of θ_1 has been given by DuMond (1957); as θ_1 increases from 0 to 90°, ψ_1 decreases from 180° to 90°. The focus occurs before the electrons return to the axis $r = 0$.

To make a spectrometer with ring focus one arranges to accept only the electrons emitted with θ's in some small range about θ_1. An annular resolving slit of radius r_1 is placed at a distance Z_1 from the source. A detector large enough to accept the beam diverging from the annular slit is placed on the axis. It is not difficult to see that electron groups with different momenta will also possess ring foci with proportionate radii and focal distances. The rings corresponding to different momenta lie on a cone with its apex at the source point. Hence varying the magnetic field strength through the nominal value for the focusing of the group under study will again produce a line.

We have not yet given any attention to the best choice of the angle θ_1, or equivalently the ratio Z_1/D. One might attempt to arrange a second-order ring focus by looking for a value of θ_1 for which both $\partial r/\partial\theta$ and $\partial^2 r/\partial\theta^2$ were zero; however, no such angle exists. The best procedure is to find the value of θ_1 for which the spectrometer accepts the largest solid angle subtended at its source for a given inhomogeneity in momentum. The calculation has been performed in slightly different ways by DuMond (1949, 1957) and by Persico (1949). The optimum value of θ_1 is in the neighbourhood of 45°, though the function maximized is sufficiently flat that any value in the range 40° to 50° is acceptable.

3.2.2. *The Pasadena ring-focus spectrometer.*

Though the power required to maintain a given magnetic field strength in a β-spectrometer is smaller if a flux path in iron almost enclosing the exciting coils is provided, there is a tendency to leave out the iron in modern precision spectrometers. The field intensity in the region of an iron-cored coil can only be found by the solution of a complicated boundary-value problem, while that for a coil immersed in an infinite, uniform, magnetic medium is readily calculated in terms of tabulated functions. Thus the iron is discarded in the interest of accurate calculation of electron trajectories.

In the instrument which we are going to describe, a detailed description of which has been given by DuMond (1957), the calculational advantages of iron-free construction are further heightened by the use of a uniform magnetic field. Forty circular coils of different radii are stacked together so that their mean turns lie on the surface of an ellipsoid of revolution of semi-major diameter 45 cm and semi-minor diameter 30 cm. The axis of revolution is the major axis of the ellipse: see figure 3.2 for a drawing of the essentials of the spectrometer. Because the coils have equal numbers of turns, and because their axial widths are equal, the arrangement provides a good approximation to an ellipsoidal current sheet with equal

currents in equal distances measured parallel to the axis. Except for the neighbourhood of the coils, the field within the ellipsoid should be remarkably uniform (Blewett 1947). The range of magnetic inductions used is 35 to 1300 gauss.

The field is measured and stabilized with the aid of three distinct proton resonance sensing elements, each designed for a different range of field strengths; because the field is uniform, a relatively large volume of sensing material can be used. Measurements of the variation of the proton resonance frequency with position in the spectrometer indicate that the departures from axial symmetry in the field are not greater than 3 parts in 10^4. Radial variations of about 0·17% have been observed in the central region of the ellipsoid, but they are not large enough to interfere seriously with the usefulness of the instrument.

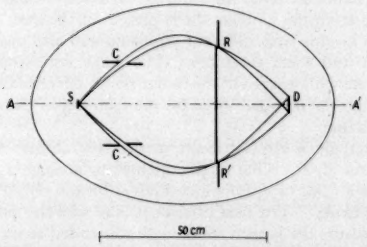


Figure 3.2. The drawing shows the trace on a plane containing the axis of symmetry of the main features of a ring-focus spectrometer. Within an ellipsoidal winding, represented by the ellipse, there is a uniform magnetic field in the direction of the spectrometer axis AA'. Electrons from the source S which are emitted at about 45° to the axis are selected by the cylindrical slit CC'. After passing through the resolving slit RR', electrons of the correct momentum enter the detector at D. The dimensions in the sketch are very close to those used in the spectrometer described by DuMond (1957). The helicoidal baffles are not shown.

To ensure that the spectrometer axis is parallel to the direction of the resultant magnetic field, it is necessary to reduce any component of the earth's field perpendicular to the spectrometer axis to a negligible magnitude. A small, non-magnetic, compressed-air driven inductor mounted near the centre of the instrument with its axle parallel to the spectrometer axis is used to measure the perpendicular component of the earth's field. The currents in two pairs of large square Helmholtz coils are then adjusted to cancel the two perpendicular components of the geomagnetic field with an accuracy to about a milligauss.

The positions of the various parts of the spectrometer within the ellipsoidal winding are shown schematically in figure 3.2. Adjustable cylindrical slit jaws enable one to select a variable range of take-off angles θ in the neighbourhood of 45°. The radius of the annular resolving slit is 17·82 cm; the detector, either a Geiger–Müller counter or a plastic scintillator, is located at 62·8 cm from the source. Adjustments of the θ-selecting slits and of the resolving slits can be made after the vacuum has been established. Not shown in the drawing are the helicoidal baffles, which reduce the number of scattered electrons arriving at the detector, and which enable one to distinguish between positrons and electrons.

To obtain a conversion-electron spectrum one inserts the source, usually a circular disc, in the source position shown in figure 3.2, and then records the electron counting rates corresponding to a set of different frequencies supplied to the radio-frequency coil of the proton resonance apparatus. The stabilization circuit adjusts the magnetic field automatically to the value corresponding to the given frequency. Provision is made for the continuous variation of the radio frequency and for the continuous recording of the counting rate.

The lower limit of the range of electron energies in which the spectrometer is useful is about 20 kev ; this bound is fixed by the field stabilization circuit. The upper limit to the energy range depends upon the amount of heat which can be generated safely in the field coils; the coils are immersed in oil, which is circulated through an external heat exchanger. The maximum electron energy at which steady operation is possible for long periods is 2·5 Mev ; for shorter periods electrons with energies up to 5 Mev can be focused.

The measured conversion-electron spectra are obtained as a function of the proton resonance frequency, which is proportional to the magnetic field in the spectrometer. To get a direct measurement of the electron momentum would require a good knowledge of the line profile. However, since the line profile does not depend appreciably upon the electron energy, at least for thin enough sources, the ratio of the momenta of two electron groups can be found very precisely by finding the number by which the abscissae of one curve must be multiplied to make that curve congruent with the other. Moreover, since the energy region in which the instrument is useful overlaps that of the crystal diffraction spectrometers, one can calibrate it against γ-rays whose energies are known with very high precision. This method has been found to be more exact than any direct calculation of the energy from the measured resonance frequency.

The geometric factors determining the line profile are the diameter of the source, the range of the angles θ accepted, and the width of the annular resolving slit. The values of the spectrometer parameters used in the design were determined by an approximate maximization of the intensity for a given line width. The thickness of the source, which also affects both intensity and line width, was not included in the optimization programme ; the inclusion of this effect would, of course, make the optimum values of spectrometer parameters energy dependent.

The best resolution at which much work has been done with the Pasadena spectrometer is about 0·1%, with a corresponding transmission calculated to be about 1% of the whole sphere. The corresponding optimum source diameter is 0·5 mm. As one example of the type of work which has been done with this spectrometer, we reproduce in our figure 3.3 the figure 37 and its legend from the paper by DuMond (1957).

3.3. *Flat Spectrometers*

3.3.1. *Principles of focusing.*

Those spectrometers in which the motion of the particles is predominantly perpendicular to the magnetic field are often termed 'flat spectrometers'. The lines of force possess a plane of symmetry and, usually, an orthogonal axis of cylindrical symmetry ; in most of the discussions the axis of symmetry will be assumed. Both source and detector are located on a circle lying in the plane of symmetry and

concentric with the axis of symmetry. We shall call this circle the optic circle, and use the symbol r_o for its radius. By adjusting the strength of the field at the optic circle H_o one can always make monoenergetic electrons of momentum p emitted in a direction tangent to the optic circle travel on that circle ; the required field strength is given by

$$H_o = p(r_o e)^{-1}. \qquad \ldots\ldots(3.11)$$

The motion of a particle which leaves the source in a direction not parallel to the optic circle depends, of course, upon the spatial variation of the field. Imagine

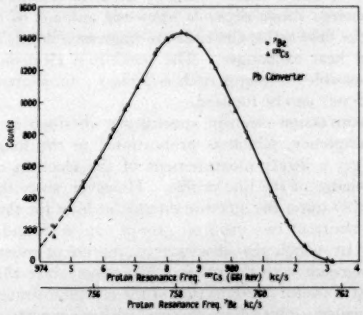


Figure 3.3. This is a reproduction of figure 37 and its legend from the paper of DuMond (1957). " Here the profile of an externally converted gamma-ray line in ^{7}Li, populated by beta decay from ^{7}Be (points indicated with circles) is superposed on the profile of an externally converted 661·52-kev gamma line from ^{137}Ba, populated by beta decay from ^{137}Cs. Both lines are formed by K photoelectrons ejected from a lead converter in precisely the same geometry. The two sets of points and proton resonance scales are here labelled in terms of the parent rather than the daughter product. The energy, $477{\cdot}3 \pm 0{\cdot}4$ kev, of the ^{7}Li gamma line determined in this way furnished an important energy calibration point for Van de Graaff studies in nuclear reactions among the light nuclei." The ^{7}Be abscissa was incorrectly drawn in the original figure ; with Professor DuMond's consent we have redrawn this scale in a manner consistent with the quoted gamma-ray energy. An energy of $661{\cdot}60 \pm 0{\cdot}14$ kev ought to have been used for the ^{137}Ba gamma-ray, though the value actually used does lie within the errors quoted.

a cylindrical coordinate system with its axis in the axis of symmetry of the magnetic system and with its origin at the centre of symmetry ; the coordinates of an electron in this frame of reference are (r, θ, Z). It is convenient to replace r and Z by the fractional radial and axial departure η and τ from the optic circle ; i.e.

$$\eta = (r - r_o)/r_o \qquad \ldots\ldots(3.12a)$$

$$\tau = Z/r_o. \qquad \ldots\ldots(3.12b)$$

Now it can be shown that, because of Maxwell's equations, the variation of the field in the neighbourhood of the optic circle is entirely determined by its variation in the plane of symmetry, i.e. by the coefficients a_n in the expansion

$$H_Z(r, Z = 0) = H_o\left[1 + \sum_{n=1}^{\infty} a_n \eta^n\right]; \qquad \ldots\ldots(3.13)$$

see, for example, Lee-Whiting and Taylor (1957).

The equations of motion, which enable us to calculate r, θ and Z as functions of time, can be transformed into a pair of orbit equations (Svartholm 1946) of the form

$$\eta'' + \omega_1^2 \eta = R(\eta, \tau, \eta', \tau') \quad \ldots\ldots(3.14a)$$

$$\tau'' + \omega_2^2 \tau = S(\eta, \tau, \eta', \tau') \quad \ldots\ldots(3.14b)$$

$$\omega_1^2 = 1 + a_1, \quad \omega_2^2 = -a_1\,; \quad \ldots\ldots(3.15)$$

the prime stands for differentiation with respect to θ. The interesting feature of the right-hand sides of equations (3.14) is that, when expanded in power series in the variables shown, they contain no terms of degree less than the second. Thus, for sufficiently small departures from the optic circle, we should obtain a good approximation to the solution by equating R and S to zero. The resulting approximate equations are very easily solved; for a point source on the optic circle one finds

$$\eta = H \sin \omega_1 \theta \quad \ldots\ldots(3.16a)$$

$$\tau = T \sin \omega_2 \theta. \quad \ldots\ldots(3.16b)$$

The constants H and T can be defined in terms of the angles fixing the direction of the initial velocity vector of the electron; note that in the first approximation H and T are the maximum radial and axial departures from the optic circle, in units of r_0.

From (3.16a) one can see that an orbit returns to $\eta = 0$ at an angle $\theta = \theta_0$ which is given by

$$\omega_1 \theta_0 \equiv (1 + a_1)^{1/2} \theta_0 = \pi, \quad \ldots\ldots(3.17)$$

if a_1 is greater than -1. The case $a_1 = 0$ corresponds to the well-known focus at 180° in a uniform magnetic field; as a_1 decreases toward -1, θ_0 increases indefinitely. If provision is made to select electrons leaving the source in the plane of symmetry at a small angle to the tangent to the optic circle, a first-order focus should be observed at $\theta = \theta_0$. If very small initial angles of inclination to the plane of symmetry are permitted, the radial focus ought not to be too much affected.

Having seen that the system will focus, we ask how it can be used as a spectrometer. To answer this question we need to know what happens to electron groups with slightly different energies. If the fractional increase in the momentum is ϵ a focus will be produced in the plane $\theta = \theta_0$ at a slightly greater radius; the increase in $\eta(\theta_0)$ is $2\omega_1^{-2}\epsilon$. Thus the instrument will function as a spectrograph, provided that the angles of emergence are sufficiently restricted that the higher-order aberrations are not too large. By placing a detector on the optic circle at $\theta = \theta_0$, and by arranging to vary the magnetic field, one can use the instrument as a spectrometer. Note that the separation of electron groups of different energies increases as a_1 is decreased toward -1.

Now let us return to the consideration of the axial motion of the electrons. We learn from (3.16b) and (3.15) that, if $a_1 > 0$, τ increases very rapidly as θ increases. If $a_1 < 0$, the orbit returns to the plane of symmetry; there is a first-order axial focus at an angle $\theta = (-a_1)^{-1/2}\pi$. As was pointed out by Svartholm and Siegbahn (1946), the radial and axial foci occur at the same angle—at $\pi\sqrt{2}$ radians—if $a_1 = -\frac{1}{2}$; we shall call spectrometers with $a_1 = -\frac{1}{2}$ the $\pi\sqrt{2}$ spectrometers. Because of the axial focus, the length of the detector remains at a practical value even when the radius of the spectrometer is considerably increased. The ability to increase the

size of the spectrometer is an advantage, since the efficiency under optimum operating conditions varies as r_0^2. Another advantage arising from the limiting of the size of the detector is the smaller effect of the natural background in the region of the spectrometer. Almost all recent spectrometers of the flat type have been built to give two-directional focusing; we shall confine our attention to this case in most of the remainder of this section.

Having decided upon the best values for a_1, we turn our attention to the higher order expansion coefficients a_n. Since the dispersion is in the radial direction, a higher-order radial focus is what is required. In general one can generate an $n+1$ order approximation to the orbits by substituting in the nth order expansion of the right-hand side of (3.14) the nth order approximation to the orbits (Svartholm 1946, Lee-Whiting and Taylor 1957). The second-order terms in the radial aberration obtained in this way are contained in

$$\eta(\pi\sqrt{2}) = \tfrac{1}{4}(1-8a_2)H^2+(\tfrac{3}{2}a_2-1)T^2; \qquad \ldots\ldots(3.18)$$

H and T are now approximations to the maximum radial and axial departures from the optic circle. Though it is possible, by suitable choice of a_2, to make either term in (3.18) vanish, the two terms do not vanish simultaneously. By choosing $a_2 = \frac{1}{8}$ one minimizes the aberrations from orbits with large radial departures from the optic circle; by choosing $a_2 = \frac{2}{3}$ the aberrations from orbits with large axial departures are minimized. In each of the two cases one would design baffles to accept more of the orbits with the less harmful kind of departure from the optic circle. The $\pi\sqrt{2}$ spectrometers are divided naturally into two classes, depending upon the choice of a_2; we shall use the adjectives 'wide-aperture' and 'high-aperture' to describe the two classes.

The field profile necessary to produce higher-order focusing for each of the two classes has been worked out. For the wide-aperture type one solves for the values of the parameters a_n necessary to cause the vanishing of the coefficients of the powers of H in the expansion of $\eta(\pi\sqrt{2})$; note that the orbits which need be considered lie wholly in the plane of symmetry, and that there is a one-to-one correspondence between the parameters a_n and the coefficients which are to be made to vanish. The required values of the field expansion coefficients have been calculated by Verster (1950) and by Pavinski (1954). For the high-aperture instrument the terms in $\eta(\pi\sqrt{2})$ which should be eliminated are those proportional to T^2, HT^2, T^4, HT^4, ...; the required values of the a_n, as far as $n = 6$, have been given by Lee-Whiting and Taylor (1957).

No compelling reason seems to exist for preferring one type of $\pi\sqrt{2}$ spectrometer over the other. If the same attention is given to fitting the field coefficients in the two cases to their optimum values, then the curves of transmission against line width should be very similar; the exact line shape will depend on the method of producing the magnetic field. There are two minor advantages in favour of the high-aperture instrument: it gives, accidentally, a perfect second-order axial focus, thus permitting the use of a shorter detector and allowing the use of a slightly longer source for the same aberration. Another advantage of the $a_2 = \frac{2}{3}$ type, which has not yet been fully exploited, stems from the fact that the field profile is very similar to $r^{-\frac{1}{2}}$. As a result the instrument has nearly the optimum values of the a_n for a range of values r_0. Calculations indicate that the definition

is good over a considerable range of electron momentum. Thus it is possible that the high-aperture spectrometer may also serve very well as a spectrograph, a feature useful in the preliminary search for conversion lines.

A discussion of flat spectrometers of a variety of types, both old and new, has been given by Siegbahn (1955). Though the high-aperture design ($a_2 = \frac{3}{8}$) has been most often used, there exists at least one example of the wide-aperture class ($a_2 = \frac{1}{8}$); this is the iron-cored instrument of mean radius 14 cm described by Zolotavin (1954). In § 3.3.2 we describe a large iron-free spectrometer of the high-aperture type.

The spectrometers we have been discussing are all in some sense descendants of the uniform-field spectrometer, with focusing at 180°, introduced by Danysz (1912). There is, however, a cognate class of instruments making use of a different type of inhomogeneous field. Instead of depending upon the distance from an axis of symmetry, the field in the plane of symmetry depends upon the distance from a line in that plane, i.e. the field in the plane of symmetry is a function of a Cartesian coordinate in that plane. The innermost orbit in the plane of symmetry is a semicircle lying entirely in a region of uniform field; the source and detector are located at the ends of this semicircle. Outside the semicircle the field falls off as a function of the distance from the line joining the source to the detector. The problem of design is to choose the field profile in such a way as to produce zero aberrations at the detector for as many orbits lying outside the bounding semicircle as possible. In the instrument described by Korsunsky *et al.* (1945) the field fell off in a number of finite steps; the size of the steps was calculated to produce zero aberration for one orbit in the region of each step. The solution to the problem of finding the field profile to produce perfect focusing of planar bundles of orbits has been given by Pavinski (1954). The Pavinski theory is the basis of the 'ketron' described by Dzhelepov and Bashilov (1950). As one would expect, this instrument is operated with a large radial aperture and a relatively small aperture perpendicular to the plane of symmetry.

3.3.2. *The Chalk River $\pi\sqrt{2}$ spectrometer.*

A large instrument belonging to the high-aperture class of two-directional focusing spectrometers has been assembled at Chalk River (Graham *et al.* 1959, private communication). The radius of the optic circle is 100 cm. The magnetic field is produced by a system of parallel, concentric, circular, air-cored coils. For each coil above a plane of symmetry there is a similar coil below that plane. In figure 3.4 (*a*) the section of the system of coils by a vertical plane containing the axis of symmetry is shown; the intersections with the optic circle are labelled C. By proper choice of the positions and turns ratios of coil pairs 1, 2, and 3 the first four field expansion parameters, a_1 to a_4, have been fitted to their optimum values. The resulting field profile is very close to $r^{-\frac{1}{2}}$ for 20 cm on each side of the optic circle. Each coil of the three pairs mentioned consists of four sections stacked one on top of the other. The windings are of aluminium ribbon insulated with mylar tape. Between the sections there are aluminium spacers containing channels for the cooling water. The coil pair labelled 4, having only one section each, provides a small region about the centre of the spectrometer wherein the spectrometer field is almost zero; the exterior field of these tiny coils has an almost negligible effect

on the focusing. All coils are connected in series. The range of field strengths at the optic circle over which the spectrometer can be used is approximately 5 to 185 gauss. The coils were manufactured and erected (by Canadair Limited of Montreal) with an accuracy to better than 1 part in 10^3.

The magnetic field is stabilized indirectly through the stabilization of the electrical current. Since there are no sufficiently large regions of sufficiently

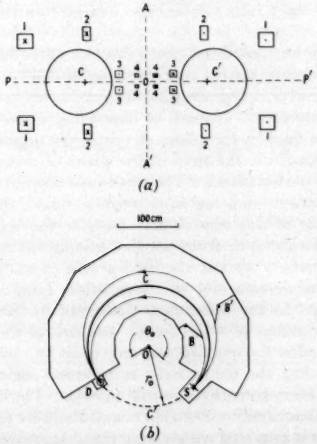


Figure 3.4. The upper part of the figure is a section of the Chalk River spectrometer by a plane containing the axis of symmetry AA′. The numbered rectangles represent the cross sections of the conductors plus adjacent cooling channels. Because the cooling channel for each coil section is below the winding, the rectangles are not quite symmetrical with respect to the plane of symmetry PP′ of the magnetic system. The optic circle is labelled CC′ in both parts of the figure; the radius r_0 is equal to 100 cm and the angle θ_0 to $\pi\sqrt{2}$ radians. The swinging baffles BB′ select the electrons emitted by the source S which are permitted to approach the detector D located behind an exit slit.

uniform field, the proton-resonance method of stabilizing the field directly is impracticable. It should be noted that the possibility of controlling the field by means of the current is one of the advantages of iron-free construction. Because of the method of field stabilization used, it is exceedingly important to control thermal expansion. By forcing water through the cooling channels between the sections of the coils, it is possible to keep the temperature of the windings within safe bounds. The vacuum chamber, to which the source and detector are attached, must also be kept from changing its dimensions; its temperature is held steady to within about 1 Fahrenheit degree by the air-conditioning of the room.

With this spectrometer it is necessary to cancel all components of the earth's field, in order to preserve the proportionality between field and current. Failure to cancel the earth's field would also affect the focusing, since the distribution of

field strength would be changed slightly. The vertical and horizontal components of the earth's field are measured with the aid of flux-gate magnetometers (Serson 1957); the sensing heads are at a sufficiently large distance from the spectrometer that the spectrometer field is negligible. The two field components are cancelled by the fields of two sets of four parallel, concentric coils; the diameter of the coils for the horizontal direction and of the larger coils for the vertical direction is about 9 metres. The calculated positions and turns ratios of these coils are such that all non-uniform terms in the central expansions of their field components of degree less than eight are absent (Lee-Whiting 1957). It is believed that the earth's field can be cancelled in the regions of the electron orbits with an accuracy to about 0·5 milligauss. A field monitor in the region of negligible spectrometer field near the centre of the instrument is used to give warning of variations in the earth's field or of possible changes in the spectrometer field.

Figure 3.4 (*b*) shows the lay-out of the spectrometer in the plane of symmetry. A thin, flat source is placed at S, with its normal tangent to the optic circle; the height of the source perpendicular to the plane of symmetry is usually much greater than the width in that plane. The detector at D is a proportional counter rather than a Geiger–Müller counter, in order to minimize dead-time losses. The baffles BB′, doors swinging on vertical hinges, allow one to vary the radial aperture; the profiles of the doors are shaped to decrease the axial aperture at small openings. When the spectrometer is in operation, the current in the coils is varied automatically in steps through the region being studied, and the number of counts at each step is recorded.

In the work done so far the instrument has been calibrated by comparison with the K-line of the 411·770 ± 0·036 kev transition in ^{198}Hg, the energy of which has been determined very accurately by diffraction of the γ-ray in crystals (Muller *et al.* 1952). The K-line of the 662 kev transition in ^{137}Ba, which follows the decay of a long-lived source, has been very precisely measured to serve as a secondary standard. Making use of the accurate value of energy of the transition in ^{198}Hg already quoted and of the electron binding energies of Cauchois (1952, 1953) as revised by Wapstra *et al.* (1959), Graham *et al.* (1959, private communication) obtained 3381·08 ± 0·30 gauss cm for the momentum of the K-conversion line of the ^{137}Ba transition; the corresponding transition energy, 661·595 ± 0·076 kev, is in good agreement with the crystal diffraction measurement, 661·60 ± 0·14 kev, of Muller *et al.* (1952). The precision with which the momenta of different electron groups can be compared is such that there is a need for more accurate standards; it is possible that these may be obtained using an electron gun as a source. Any attempt to calibrate the spectrometer directly by calculating the field from the distribution of current would be frustrated by the difficulty in measuring the radius r_0 with precision.

The spectrometer design is believed to permit steady operation at the currents necessary to focus electrons of energy as high as 5 Mev; changes in coil dimensions should be less than 3 parts in 10^5. Measurements of the resistance of the coils enable one to calculate that the thermal expansion is less than 1 part in 10^5 when 3 Mev electrons are being focused. The lower limit to the useful range of electron energies is fixed by the thickness of the counter window; at present electrons with energies as low as 20 kev can enter the counter.

As an example of the work which can be done with this spectrometer we show in figure 3.5 some of the conversion lines of the 411 kev transition in ^{198}Hg, measured with a resolution of about 0·015% and a transmission estimated to be about 0·1% (Graham *et al.* 1959, private communication). Note that the K-line is appreciably wider than the others; the difference in width is exp'ained by the

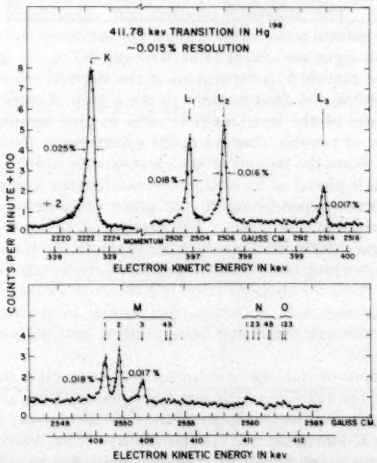


Figure 3.5. Some of the conversions lines of the 411 kev transition in ^{198}Hg measured with the Chalk River $\pi\sqrt{2}$ spectrometer at a resolution of about 0·015% and a transmission of about 0·1% (Graham *et al.* 1959, private communication).

greater intrinsic width of a vacancy in the K-shell. The width of the source used in the measurements was about 0·015 cm, the length 1·3 cm, and the thickness about 50 μg per cm^2. A calculation of the radial aberration caused by the finite source height indicates that the length of source used was far below the optimum value; indeed, an increase of this length to 4 cm should have little effect on the line width.

3.4. *Conclusion*

In this subsection we shall compare a few of the properties of ring-focus and $\pi\sqrt{2}$ spectrometers, using the two instruments already described as examples. Our aim is to assist the reader to decide which type of instrument would be most suitable for a particular task.

Precision is possible with both the instruments described, because an accurate knowledge of the electron orbits was available to the designer. This knowledge, it should be pointed out, depends on our ability to calculate with great accuracy the fields of systems of currents in an infinite medium of uniform permeability, i.e. in the absence of iron. For the Pasadena spectrometer there exist exceedingly

simple analytical formulae for the orbits. For the Chalk River instrument the analytical formulae for orbits are approximate, and are much more complicated; however, the adequacy of these formulae has been checked by numerical integration. Note that another great advantage of iron-free construction is the possibility of strict proportionality between field strength and current.

We shall compare the efficiency of operation of our two examples at a resolution of 0·1%. As already stated, the transmission of the Pasadena spectrometer is about 1% at this resolution and the optimum source area is $\frac{1}{4}\pi(0{\cdot}05)^2 \simeq 0{\cdot}002$ cm². Measurements have been made with the Chalk River instrument at this resolution with a source area of $0{\cdot}15 \times 2{\cdot}5 \simeq 0{\cdot}37$ cm²; the transmissions were estimated to be 0·3%. Thus, if the amount of source material available is just about enough to make a source for the Pasadena spectrometer, that instrument is to be preferred because of its threefold larger transmission. If large quantities of source material are available, then the Chalk River spectrometer is superior, because its luminosity (the product of transmission and source area) is about 60 times greater. Since size of source area depends upon the size of the instrument, the standing of the ring-focus type can be improved by making a larger spectrometer; indeed, if equal cost of construction were the criterion, it would be necessary to magnify the Pasadena machine before making the comparison. However, since the length of the Chalk River source could be increased four or five times without spoiling the resolution (still at 0·1%), the conclusion would not be changed; the instrument described by DuMond operates already with the optimum ratios of dimensions.

For both machines the useful energy range is about 20 kev to 5 Mev; the upper limit depends upon the rate of removal of heat. Warming up is much more undesirable for the Chalk River instrument, because it affects the proportionality constant between momentum and current, the spectrometric variable. For the same reason an accurate cancellation of the earth's field is also more important in the Chalk River instrument.

Nothing has been said so far about background, because few general statements seem to be possible. In both cases there may be, of course, a continuous background of nuclear γ-rays. Though no measurements or calculations are available for comparison, the background of scattered electrons getting through the exit slit is usually assumed to be smaller for a lens type spectrometer, because of its helicoidal baffles. The general background of penetrating radiations from the environment should be smaller for the $\pi\sqrt{2}$ instrument because its detector, located just behind the focus, should be smaller; this comparison applies only to instruments of comparable size.

Calculations of line shape have been carried out for the Pasadena machine; analogous calculations and comparison with experiment for a wide-aperture $\pi\sqrt{2}$ spectrometer have been described by Zolotavin and Sadovski (1954). In each case the calculation was confined to geometric effects, the energy loss in the source being ignored. Generally, workers with conversion electrons have sought precision by striving for high resolution rather than for a very accurate knowledge of the line shape.

A practical advantage of most $\pi\sqrt{2}$ spectrometers over the Pasadena design of the ring-focus spectrometer is the greater accessibility to the operator of the source and the detector: compare figures 3.2 and 3.4.

§ 4. Spectrometers using the Compton Effect

4.1. *Introduction*

The Compton effect is the elastic collision of one photon and one electron. In the applications to be discussed the electron is in one of the orbits of an atom of low atomic number. We shall assume, as is usually done, that the electron is free and that it is at rest before the collision. These are not unreasonable assumptions, since the electron binding energies are very small compared with the energies of the photons considered. Quantitative estimates of the effects of the initial motion of the electrons will be discussed later.

The initial energy of the photon and the final kinetic energy of the electron, both in units of the electron rest energy, will be represented by α and E respectively. Let the angle of ejection of the electron with respect to the incident photon direction be ϕ, and let the angle of scattering of the γ-ray be θ. From the conservation of energy and momentum one can obtain (see, for example, Davisson and Evans 1952) the result

$$E = E(\alpha, \phi) \equiv 2\alpha^2[1 + 2\alpha + (1+\alpha)^2 \tan^2\phi]^{-1}. \qquad \ldots\ldots(4.1)$$

Thus the energy of the electron after the collision is a function of two variables, the photon energy and the angle of ejection of the electron. On the other hand, if E and ϕ are measured, the value of α can be determined by solving the quadratic equation obtained from (4.1). If $E\tan^2\phi < 2$, a condition which inspection of equation (4.1) will show to be valid for the Compton process, there exists one and only one positive solution to this quadratic. Thus an apparatus which measures the quantities E and ϕ for an electron set in motion by the Compton process will serve as a spectrometer for γ-rays.

Alternatively, we can also determine the energy of a photon by measuring the electron energy and the angle of scattering of the photon. One can easily show that there is a simple relation between θ and ϕ for fixed α (it is convenient to use the variable $\theta' \equiv \pi - \theta$):

$$\tan\tfrac{1}{2}\theta' = (1+\alpha)\tan\phi. \qquad \ldots\ldots(4.2)$$

Substitution of (4.2) in (4.1) also gives rise to a quadratic equation in α. This equation has a unique positive solution for all permissible values of E and θ'. Thus the values of E and θ' determine the value of α.

Now suppose that we allow a beam of monoenergetic γ-rays to fall upon a piece of material, to be called the radiator, of low atomic number. If the binding of the electrons is sufficiently weak compared with the photon energy, then the theory of the Compton effect which we have been discussing will hold approximately. Note that the radiator must be thin enough that the electrons can emerge without excessive energy loss. From our earlier discussion it follows that any apparatus which measures the electron energy E and one of the angles ϕ or θ can be used as a Compton-electron spectrometer for γ-rays. In a precision instrument it is probably always advisable to determine the electron energy by means of a magnetic spectrometer; see § 3. In the spectrometers we are going to consider the measurement of the angle is made by selecting electrons or photons which pass through a fixed system of apertures. For such a selection to be possible it is necessary that the variation of the directions of the incident photons at any point on the radiator be small. In principle one should use a point source or a

parallel beam; in practice the size of the source or the divergence of the beam, as well as the angle-selecting aperture, are of finite but tolerable magnitude.

In § 4.2 we shall discuss the development of Compton spectrometers for use in the study of γ-rays. General properties of some Compton spectrometers are discussed in § 4.3. As will become evident, the efficiencies of the Compton spectrometers which have been described in the literature are somewhat inferior to those of recent spectrometers based on other γ-interactions. Since the Compton type of device appears to have been developed less intensively than the other types, we believe that the inclusion in § 4.4 of two proposals for improvement will give a fairer picture.

Before proceeding to the description of actual instruments, we shall give some attention to the problem of deciding upon the best value of ϕ to use in constructing a spectrometer. For a given photon energy and a given value of the minimum accepted value of ϕ, say ϕ_1, the range of accepted values of ϕ, $\chi \equiv \phi_2 - \phi_1$, is fixed by the desired resolution. Our problem is to find the value of ϕ_1 which gives the greatest efficiency; because of the relation (4.2) there is no need to treat the θ-selecting case separately. First we shall investigate how the solid angle lying between the cones $\phi = \phi_1$ and $\phi = \phi_2$ depends upon the value of ϕ_1. In order to do this, assume that a γ-ray with photon energy $\alpha_2 = \alpha(1-r)$ may just be resolved from the ray of energy α, if ϕ_2 is so chosen that the electrons ejected by the photon of energy α in the direction $\phi = \phi_2$ have the same energy as those ejected by the photon of energy α_2 at $\phi = \phi_1$; we assume here that r is positive. In terms of the function defined in (4.1) the equation for ϕ_2 is

$$E(\alpha, \phi_2) = E(\alpha_2, \phi_1). \qquad \ldots\ldots(4.3)$$

An approximate solution of (4.3), obtained by retaining only the first-order terms in r and χ, is

$$\chi \simeq [2r(1+\alpha)^{-1}]^{1/2}, \quad \text{for} \quad \phi_1 = 0; \qquad \ldots\ldots(4.4a)$$

$$\chi \simeq r(1+\alpha)^{-1}\cot\phi_1, \quad \text{for} \quad \chi \ll \phi_1. \qquad \ldots\ldots(4.4b)$$

Let Ω be the solid angle defined by the cones at ϕ_1 and ϕ_2. Then

$$\Omega = 2\pi(\cos\phi_1 - \cos\phi_2) = 4\pi \sin(\phi_1 + \tfrac{1}{2}\chi)\sin(\tfrac{1}{2}\chi). \qquad \ldots\ldots(4.5)$$

Substitution of either (4.4 *a*) or (4.4 *b*) in (4.5) leads to

$$\Omega \simeq 2\pi r(1+\alpha)^{-1}\cos\phi_1. \qquad \ldots\ldots(4.6)$$

Not unexpectedly Ω decreases as the resolution is improved, i.e. as r is decreased, and also as the photon energy is increased. More surprisingly, formula (4.6) shows Ω to be a monotonic decreasing function of ϕ_1. This conclusion is confirmed by a more precise solution of equation (4.3).

It is known that the differential cross section for the ejection of a Compton electron decreases rapidly as the angle between the scattered electron and the incident photon is increased, see Davisson and Evans (1952) or Nelms (1953). Because of the behaviour of the cross section, and because of the falling off of utilizable solid angle, one can say that the highest counting rate at a given resolution should be achieved with $\phi_1 = 0$, though the decrease in efficiency is very slight for small values of ϕ_1. The choice of the type of magnetic spectrometer (see § 3) to be employed depends to some extent upon the value of ϕ_1 to be used. Thus a lens-type spectrometer cannot be used with $\phi_1 = 0$. Indeed, all the instruments that we shall

discuss are of the flat type. One reason for the predominance of flat spectrometers is that means have been found to use them with broad radiators, an improvement not yet effected for the lens type. Some Compton spectrometry has been done with lens spectrometers, see for example Motz (1956), but there is no indication that this type of instrument is particularly suited for high precision. The Motz arrangement differs from those described in § 4.2 in using a coincidence with the back-scattered photon to reduce the background. Thomas and Lauritsen (1952) have used a lens-type Compton spectrometer in the study of γ-rays from reactions in light nuclei; intensity measurements were made with a method involving the use of a thick converter.

Having decided to use $\phi_1 = 0$ we get a simplification of formula (4.1):

$$E = 2\alpha^2(1+2\alpha)^{-1}. \qquad \ldots\ldots(4.7)$$

The E in (4.7) corresponds, of course, to the high-energy edge of the line. The problem of calculating α from a known E is also simplified:

$$\alpha = \tfrac{1}{2}[E+(E^2+2E)^{1/2}]. \qquad \ldots\ldots(4.8)$$

At the beginning of the discussion of Compton spectrometers we made the simplifying assumption that the electrons in the radiator were free and stationary.

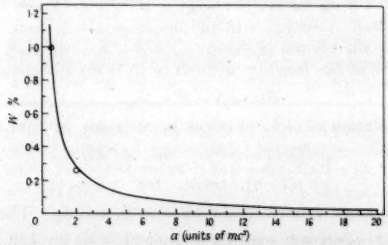


Figure 4.1. The intrinsic half width of the electron momentum line as broadened by the initial motion of the electrons in a nitrogen atom is plotted against photon energy in units of m_0c^2. The circles are the points calculated by Dzhelepov (1955).

Since these electrons are in reality bound in atoms and have a distribution of momenta, the momentum line of electrons ejected in the forward direction by photons of a given energy will have a finite width. DuMond (1933) has shown that the initial electron motion can be used to explain the energy width of the scattered γ-ray; calculations of the shape of the electron line have been performed by Dzhelepov (1955) and by Lee-Whiting (1960). As an approximation to an average atom in 'Cellophane', the nitrogen atom was used; the distribution of momenta was calculated from some approximation to the spatial wave functions. One finds that it is the momentum rather than the energy of the target electron which is important. In figure 4.1 we show Lee-Whiting's curve for the intrinsic momentum line width of a Compton electron spectrometer. Note that this width rises very rapidly as the photon energy goes below 1 MeV; for larger photon energies it is of the order of 0·1%.

4.2. *Development of the Flat Compton Spectrometer*

The simplest form of flat Compton spectrometer is illustrated by figure 4.2 (*a*). From it are descended all the other instruments which we shall describe briefly in this subsection. Gamma rays from a source S are allowed to impinge on a thin target R (to be called the radiator) which replaces the entrance slit of a uniform-field 180° β-spectrometer. The detector lies behind the exit slit of the spectrometer.

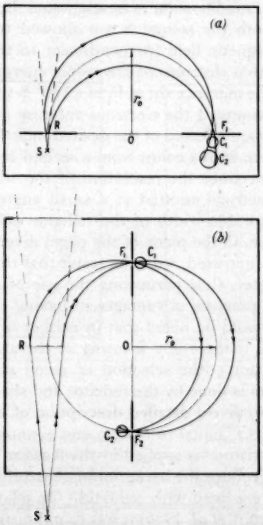


Figure 4.2. In the upper section we illustrate the principle of the Compton spectrometer with a narrow radiator as used by Latyshev and his co-workers ; the source is labelled S, the radiator R, and the counters C_1 and C_2. In the lower section the broad radiator type, suggested by Dzhelepov and Orbeli (1948) and used in the ' ritron ', is shown.

The range of values of ϕ, the angle between the incident γ-ray and the recoil electron, in this sort of instrument depends upon the ratios of the source and radiator dimensions to the distance SR. The precision of the determination of the value of the electron momentum depends, of course, on the aberrations of the β-spectrometer. The resolution of the Compton spectrometer also depends significantly upon the change in direction and energy suffered by the electron in leaving the radiator and upon the scattering of the electrons by the atoms of any gas left in the instrument.

The first use of a spectrometer of the type of figure 4.2 (*a*) appears to have been made by Latyshev and his collaborators (see Latyshev 1947) who achieved a resolution to about 2% working with 500 mc sources of small dimensions. The

background was much reduced by the use of a coincidence telescope of Geiger–Müller counters as the detector.

One method of improving the Latyshev type of instrument is to reduce the uncertainty in the measurement of the electron momentum by using a better form of β-spectrometer. Mladjenovic and Hedgran (1952) substituted a $\pi\sqrt{2}$ spectrometer for the 180° type used by Latyshev. Wedge type spectrometers have been used by Motz *et al.* (1953) and by Ulmer and Ziegler (1956).

A more fundamental modification was suggested by Dzhelepov and Orbeli (1948). Gamma rays from the source S are allowed to fall upon an extended radiator in a uniform magnetic field (perpendicular to the page in figure 4.2(*b*)). The radiator is so placed that electrons of the proper energy recoiling in the forward direction are focused on the entrance slit at F_1 of a 180° β-spectrometer, after turning through about 90°. By counting the electrons entering the counter C_2 behind the second slit at F_2, particles are selected of the desired momentum ; by retaining only those counts in coincidence with a count from a second counter C_1 the background is greatly reduced. Now, since the resolution of the whole instrument depends upon the selection of electrons emitted at a small angle to the incident gamma direction, the greater allowable width of the radiator does not permit the use of beams of greater radial (i.e. in the plane of the page) divergence at S. Instead the distance R to S is greatly increased, with the result that the radial dimension of the source can be much greater, thus permitting the use of sources of lower specific activity. There are also practical advantages in having the source moved farther from the counters. It should be noted that in neither of the two types of instrument shown in figure 4.2 is there any focusing in the axial (perpendicular to the plane of the page) direction ; the selection of recoil electrons with small axial components of momentum is done by the radiator and slit heights.

Dzhelepov *et al.* (1954) give a detailed description of an instrument of the type of figure 4.2(*b*), built with r_0 equal to 5·5 cm and a source to radiator distance of about 31 cm ; the name ' ritron ' is applied by the Russian physicists to instruments of this type. In order to reduce the background of secondary electrons the interior walls of the instrument were lined with celluloid. In a later and improved version of the instrument (Dzhelepov *et al.* 1956) it was found possible to attain a resolution to about 3·5% at an efficiency of $8{\cdot}7 \times 10^{-7}$ peak coincidences per photon for gamma energies of about 1 Mev. The resolution deteriorates rapidly at lower energies in all instruments of this type, because of multiple scattering in the radiator.

Improved focusing in instruments descended from the ' ritron ' has been effected in two ways—by shaping the radiator and by using an inhomogeneous magnetic field. Dzhelepov (1951) has shown that a surface exists such that the path of a forward-recoiling electron produced at any point on the surface, by a photon from a point source, passes through a given straight line ; it is assumed that the surface and the given straight line both lie in a uniform magnetic field. Thus, if we could make the radiator in the shape of a section of this surface, there would be no aberration at the first slit. Furthermore, it would no longer be necessary to demand that the angles subtended by the radiator at the source be small ; the range of angles ϕ to be accepted would be fixed by the width of the first slit. Because the surface discussed by Dzhelepov cannot be rolled out upon a plane, it would be exceedingly difficult to construct one of thin ' Cellophane '. In

practice it has been customary to use a right cylinder erected upon the line of intersection of the Dzhelepov surface and the plane of symmetry of the instrument. Limitation of ϕ in the axial direction is effected by the slit heights as in the earlier instrument.

The ' elotron ', described by Dzhelepov and Shestopalova (1956), is one of the instruments that have been constructed to operate with a curved radiator. In figure 4.3 the radiator is labelled R. This spectrometer has a mean radius of about 6·5 cm and a source to radiator distance of about 50 cm. The ' elotron ' has also

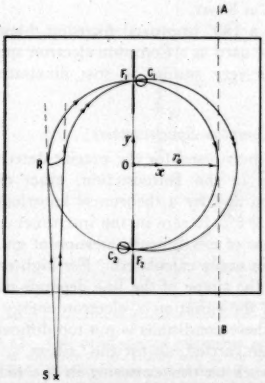


Figure 4.3. This sketch illustrates the focusing in the ' elotron '. The radiator R is curved to provide an improved focus at F_1 for the forward-ejected electrons. To the right of the line AB the field falls off as a function of the Cartesian coordinate x in such a way as to improve the focus at F_2.

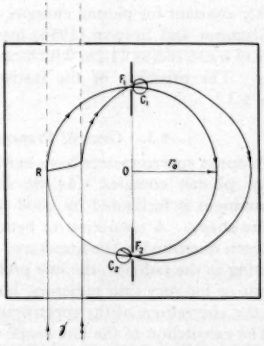


Figure 4.4. In the instrument described by Groshev *et al.* (1956) a collimated beam of γ-rays from a large source in a reactor ejects electrons from a curved radiator R. The forward-ejected electrons pass through a slit at F_1. Outside the circle of radius r_0 the field falls off as a function of distance from the point O in such a way as to produce an improved focus at F_2.

some improved focusing between the two counters. The magnetic field is uniform in the region to the left of the line AB. To the right of AB the field falls off as a function of x in the manner suggested by Pavinski (1954), see § 3.3 ; there is also some undesired dependence on y, but its effects on the focus are believed to be unimportant.

Groshev *et al.* (1956) have constructed a spectrometer in which the focusing between the two slits is obtained by means of a magnetic field which falls off radially. The principle of the instrument is illustrated by figure 4.4. This spectrometer has been used extensively in the study of neutron-capture γ-rays produced in a sample placed in the thermal column of a reactor. Because the breadth of the γ-source is not small compared with the radius of the spectrometer it is necessary to collimate the beam, i.e. to create the illusion of a point source at infinity. Neutron capture γ-rays were collimated to about 1° and allowed to fall upon the curved radiator R.

The radiator lies just within the central region (of radius $r_0 = 6$ inches), in which the magnetic field is uniform. The radiator is so curved that monoenergetic electrons ejected in the forward direction are brought to a focus at F_1 when the magnetic field is correctly adjusted for the energy of the electrons. After passing through the slit at F_1 the electrons are in the region outside the circle of radius r_0. Here the magnetic field falls off radially in such a way as to produce an improved focus at F_2. The instrument gives $2{\cdot}5 \times 10^{-7}$ peak coincidences per photon at a resolution of 2·5% in the regions of 1·33 MeV. It is particularly noteworthy for its results at higher energies; the measured resolution has been shown to remain roughly constant for photon energies as high as 8 MeV.

Glazunov and Spitsyn (1956) have built a 180° improved focusing β-spectrometer of mean radius 11 cm which can also be used as a Compton electron spectrometer. The principle of the instrument is very similar to that illustrated by figure 4.3.

4.3. *General Properties of Compton Spectrometers*

Compton spectrometers have not usually been used for the precise determination of photon energies. As we suggested in the Introduction, exact energy measurement is facilitated by good resolution and by a theoretical knowledge of the line shape. A resolution to better than 2·5% is rare in the instruments that have been described in the literature. Because of the difficult question of multiple scattering in the radiator, the line profile is not easily calculated. For high-energy photons or for very thin radiators, however, the shape of the line depends mainly upon the aberrations of the spectrometer and the variation of electron energy with ϕ. The calculation of the line shape under these conditions is not too difficult for a uniform-field machine, and has actually been carried out for the 'ritron'. The results of the calculations, including some work on the scattering in the radiator and in the residual gas in the spectrometer chamber, are summarized by Dzhelepov *et al.* (1954). Unfortunately the calculated curves do not fit the experimental ones with sufficient accuracy to be of much use in precision energy measurements. Calibration of Compton spectrometers has usually been effected with the aid of γ-rays whose energies have been determined accurately by other means, rather than by direct measurement of field strength and r_0.

In contrast to its performance in measuring energies, the Compton spectrometer is particularly well suited to determining the intensities of γ-rays. The Klein–Nishina formula for the cross section for the Compton effect is very reliable compared with the approximate formulae available for the photoelectric effect and for pair production. Calculations of the spectral sensitivity, i.e. of the area under the line, as a function of energy have been carried out for the 'ritron' (Dzhelepov *et al.* 1954) and for the spectrometer of Groshev *et al.*, see figure 4.5. The calculated curves have been checked by comparison with measurements of cascade γ-rays from ^{60}Co and ^{24}Na, for which the relative intensities are known. The relatively slow variation with energy of the curves in figure 4.5 is another advantage of Compton spectrometers for intensity measurements.

A discussion of the resolution of the 'ritron' under various conditions has also been given by Dzhelepov *et al.* (1954, 1956). In figure 4.6 we show the curves of measured resolution against photon energy for the improved 'ritron', for the

' elotron ' (Dzhelepov and Shestopalova 1956), and for the machine of Groshev *et al.* (1956). Corresponding figures for the transmissions, or the numbers of peak coincident counts per photon incident upon the radiator, are available for two of the spectrometers. The ' ritron ' yields $8{\cdot}7 \times 10^{-7}$ peak coincidences per incident

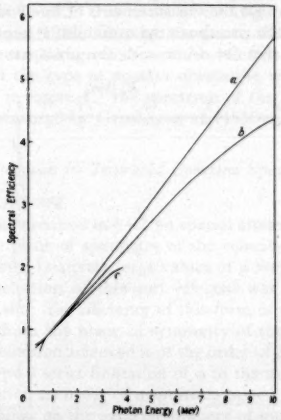


Figure 4.5. Calculated spectral efficiencies, or areas under the line, are shown, *b*, for the spectrometer of Groshev *et al.* (1956) and *c*, for the ' ritron ' (Dzhelepov *et al.* 1954). Curve *a* is the Klein–Nishina cross section for the ejection of an electron in the forward direction. All the curves are normalized to unity at 1 MeV.

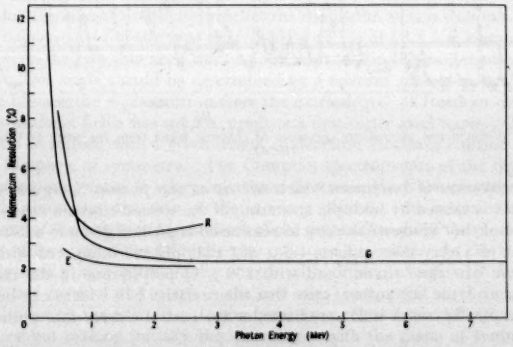


Figure 4.6. Measured resolutions are plotted against photon energy : R, for the ' ritron ' (Dzhelepov *et al.* 1954), E for the ' elotron ' (Dzhelepov and Shestopalova 1956), and G for the instrument described by Groshev *et al.* (1956).

1·022 MeV photon; this number was computed from data supplied in the two references mentioned in the first sentence of this paragraph. Groshev *et al.* state that there are $2{\cdot}5 \times 10^{-7}$ peak coincidences per 1·33 MeV photon with their instrument.

As the energy is increased the resolution approaches a more or less constant value, attributable to the geometric properties of the instrument (Dzhelepov *et al.* 1954). It is not certain that the measurements of line widths are good enough to enable one to state that the rise above the minimum at about 1 MeV in the ' ritron ' curve is real; this rise did not occur with the unimproved ' ritron '. The very

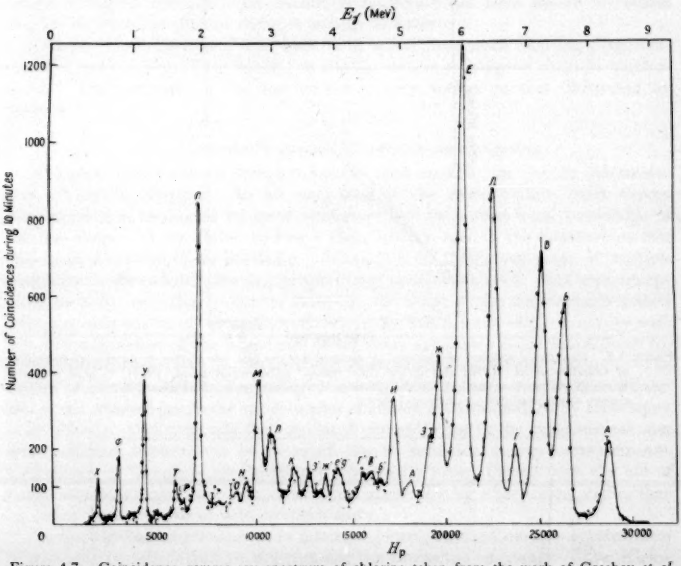


Figure 4.7. Coincidence gamma-ray spectrum of chlorine taken from the work of Groshev *et al.* (1956).

rapid deterioration of resolution which occurs as the photon energy is decreased below 1 MeV is caused by multiple scattering of the ejected electron in the radiator. In the case of the ' elotron ' the rise in resolution is pushed down to a lower energy by the use of a very thin radiator—5 μ of ' Cellophane ' compared with 50 μ of ' Cellophane ' in the ' ritron ' and with 100 μ of polystyrene in the machine of Groshev *et al.*; the last authors state that their relative half widths can be reduced to 2·3% at 500 kev with a 25 μ radiator, at the cost of a three-fold diminution of intensity.

The background counting rate has been maintained at a low level in the instruments under discussion by the requirement of coincidence between a pair of

counters placed one behind each slit. Measurements of the background were made by Groshev *et al.* (1956) by replacing the sample in the reactor by a slightly smaller one coated with a thin layer of boron carbide. The measured background, which is practically independent of the sample material, rises slowly to a maximum of about five or six coincidences per minute at the magnetic field setting corresponding to a photon of about 2 MEV. One result of the improvements effected on the 'ritron' between 1954 and 1956, whereby the gas in the chamber was removed, was a great reduction in the background caused by Compton ejection in the gas.

As an illustration of the type of spectra obtainable with a Compton electron spectrometer, we show in figure 4.7 the spectrum of the neutron-capture gamma rays from chlorine as measured by Groshev *et al.* (1956).

4.4. *Proposals for Improved Compton Spectrometers*

4.4.1. *Two-directional focusing.*

In the spectrometers described in § 4.2 no special attempt was made to limit the initial inclination to the plane of symmetry of the velocity vectors of the electrons accepted; in other words, relatively large values of ϕ were permitted in the axial direction. The only selection of this sort effected was that resulting from the finite height of the exit slit; the efficiency of this form of selection varies considerably with the distance from the plane of symmetry of the point of ejection of the electron. When the resolution achieved is of the order of 3%, no great harm comes from the failure to ensure a strict limitation of ϕ in the axial direction. However, if one aims for a resolution an order of magnitude better, it is desirable to arrange for a more precise selection on the axial component of momentum of the accepted electrons.

The problem of adding the appropriate form of axial focusing to a spectrometer of the type described by Groshev *et al.* (1956) has been considered by one of the authors of this report. Though the principle involved is more generally applicable, we shall, for the sake of simplicity, restrict the discussion to this example. A non-uniform magnetic field of the type described by (3.13) of § 3.3.1 is assumed. The angle between the two slits need not be 180°, as it was in the earlier machines, see figure 4.8; this angle should be determined by a relation analogous to (3.17).

In the case of the $\pi\sqrt{2}$ spectrometers the extra degree of freedom obtained by using non-uniform fields was used to produce a first-order axial focus. This focus enables one to collect, with a given length of detector, electrons emitted at steeper angles to the plane of symmetry. For Compton spectrometers of the type that we are discussing one wishes to select only those electrons which are emitted in a small solid angle about the direction of the incident photon; the selection should be independent of the height in the radiator of the point of origin of the electron. Thus the forms of focusing desired for the two cases are opposites. In the former case orbits originating in a point in the source with different angles to the plane of symmetry are focused at a point in the exit slit; each point in the source has a different image on the exit slit. In the latter case orbits from different heights in the radiator but making initially the same angle with the plane of symmetry pass through one point on the exit slit; orbits starting out at different angles to the plane of symmetry pass through different points in the plane of the exit slit.

Approximate equations determining the required angular separations of the radiator and slits in figure 4.8 have been set up and solved by Lee-Whiting (1960). A solution can be obtained, in principle, for any value of the field expansion parameter a_1 in the range $-0{\cdot}1$ to $-0{\cdot}2$. When higher-order aberrations and some other effects are taken into account a value of a_1 in the neighbourhood of $-0{\cdot}14$ is suggested.

In contrast to figure 4.4, the arrangement of figure 4.8 has the electrons moving in a non-uniform field in the region between the radiator and the first slit. Thus the calculation of the shape of the radiator needed to ensure that the forward-ejected electrons of the correct energy get through the first slit is more complicated.

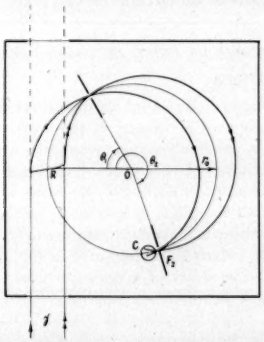


Figure 4.8. The motion of the electrons in the type of spectrometer proposed by Lee-Whiting (1960) is sketched. The field is a function of distance from the point O.

It can be carried out, however, both for a parallel beam of incident γ-rays and for a point source, if the first few expansion coefficients a_n are known. Note that if the radiator is placed roughly symmetrically astride the optic circle, instead of all on one side of it as in figure 4.4, there should be a smaller line width for the same counting rate.

If one attempts to build a Compton spectrometer for greatly improved resolution, there will eventually be difficulty from the interaction of the electrons with the windows and gas of the counter shown behind the first slit in figures 4.2(*b*), 4.3 and 4.4. A counter can probably be constructed to give an energy spread sufficiently narrow to be tolerable for use at a resolution of about 0·3% with the electrons corresponding to a photon of energy 1 MeV. The effect on the axial focusing of the change in directions caused by multiple scattering is not so easily dismissed. We have left the first counter out of figure 4.6, because it may be necessary to remove it for operation at low energy. Since the amount of multiple scattering decreases rapidly as the electron energy is increased, the two-counter mode of operation should certainly be possible for photon energies greater than two or three MeV. In removing the first counter from the design one would destroy one of the great advantages of the Compton spectrometer—its low background counting-rate.

Alternative methods of reducing the background are to use coincidences with the scattered photon (see §4.4.2), to use a counter telescope behind the final slit as in figure 4.2 (*a*), or to increase the possibilities for shielding by making the instrument bigger. An advantage, in the quest for low background, of the instrument that we have been considering is the relatively small size of its exit aperture ; indeed, in the first-order theory of orbits this aperture reduces to a point. A smaller detector, less influenced by environmental background, can be used with a smaller exit aperture.

Unfortunately no detailed experimental results are available for a Compton spectrometer operating with a resolution better than 1%. In order to be able to make some sort of comparison with other types of γ-ray spectrometers operating

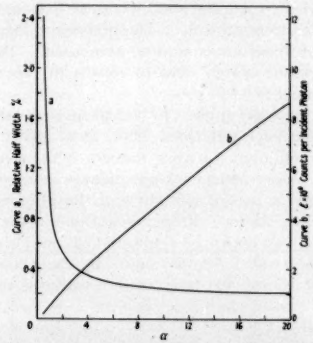


Figure 4.9. The variations with photon energy of, *a*, the estimated line half width, and, *b*, peak efficiency, for the proposed spectrometer of Lee-Whiting (1960).

in the ' precise ' range of resolutions, we shall quote the estimates of efficiency and resolution for the design of Lee-Whiting (1960). The reader is warned that these estimates depend on rather gross approximations, and that they have not been checked by experiment ; in addition, the measures which might be needed to reduce the background could have a deleterious effect on the efficiency. Curves of resolution and of efficiency (in peak counts per photon) against the photon energy are given in figure 4.9 for an arrangement optimized for 0·35% resolution at about 2 MeV. The effect on the resolution of the line broadening caused by initial electron motion is included. The rapid deterioration of resolution below 1 MeV is caused mainly by multiple scattering of the electrons in the radiator. The monotonic increasing nature of the graph of efficiency in figure 4.9 follows from the similar behaviour of the differential cross section for the ejection of an electron in the forward direction. In the derivation of the expression for efficiency it was possible to assume that the angular width of the forward lobe of the differential Klein–Nishina cross section was much greater than angular dimensions of the acceptance solid angle of the spectrometer. At higher energies this assumption is no longer valid. Thus, because the total

Compton cross section falls off for increasing α, the efficiency must eventually become a decreasing function of α.

4.4.2. *Use of the back-scattered photon.*

By working with coincidences between the forward-ejected electron and the back-scattered photon, one can hope to reduce the background counting-rate without introducing the energy loss associated with an intermediate electron counter. The angle between the back-scattered photon direction and the reverse of the incident photon direction θ' is always greater than the angle between the ejected electron direction and the incident photon direction ϕ. Formula (4.2) shows that, if ϕ is small, θ' is many times larger than ϕ, even for a photon energy as low as 1 MeV. If θ' is large enough in an appreciable fraction of the collisions under consideration, then the corresponding back-scattered photons can be detected without interference from the much stronger incident beam. The practical difficulties would be much lessened if a γ-beam of small cross section were used. Thus it is desirable to abandon the broad radiator design, and to return to the narrow radiator type considered at the beginning of § 4.2.

A preliminary report of what appears to be a promising use of the method of the back-scattered photon has been published (Motz *et al.* 1959). A collimated beam of γ-rays was allowed to fall upon a narrow radiator in the source position of a $\pi\sqrt{2}$ spectrometer of mean radius 34 cm. Coincidences were recorded between the electrons passing through a curved exit slit and the photons observed in a NaI detector near the primary beam. Proper selection of the angle of the back-scattered photon results in an improved resolution. Lines have been studied with a resolution of 0·5% in the range 0·5 to 6·5 MeV. By using four parallel exit slits in the focal plane, Motz *et al.* are able to increase greatly the rate at which data can be taken.

§ 5. Spectrometers using the Pair Effect

5.1. *Introduction*

Of the various interactions we have considered as bases for the detector of γ-rays, only pair production has a total cross section which increases with energy. It follows that at sufficiently high energies, spectrometers which depend upon pair production may be expected to become more efficient at a given resolution than those based on any of the other three processes. Magnetic pair spectrometers have indeed been used for the measurement of γ-rays of energy greater than 100 MeV. For an example of the precision (to 0·1%) that may be achieved at these energies, the reader is referred to the work of Crowe and Phillips (1954). The present discussion will be limited to instruments used in the study of lower energy reactions in which the γ-ray energies encountered are usually less than, say, 20 MeV. It is to be noted that pair production has zero cross section over an appreciable fraction of the energy range of interest and therefore cannot, even in principle, be used as the basis for a universal spectrometer. On the other hand, the opportunity for discrimination against background by coincidence techniques using the electron and positron, or one or both of these combined with the annihilation quanta, makes the pair process attractive as a method for γ-ray detection in spite of the energy limitations. In

this section we shall attempt to assess the precision, transmission, and other performance characteristics of different types of pair spectrometer.

The experiments in which pair spectrometers are used as γ-ray detectors tend to fall into two categories depending on the properties of the γ-ray source. The source may be either a broad source such as may be produced in a reactor by neutron capture in an extended target, or a 'point' source such as that formed when a target is bombarded by a narrow collimated beam of particles. Gamma rays from radioactive materials would also normally be studied as point sources, but, because of their limitation to energies greater than 1·02 MeV, pair spectrometers have only a minor application in this field. For detecting γ-rays following accelerator induced reactions it is usually necessary to work with converters of small size situated very close (within a few inches at most) to the target, or in effect, identical with it if the internal pair conversion process is used. Therefore, for experiments of this type, it is possible to use various forms of lens-type (§ 5.2) pair spectrometers which require sources or converters of small area in order to give good resolution. On the other hand, to measure the near-parallel γ-ray beams emitted by distant sources such as those encountered in neutron capture γ-ray experiments, it is advantageous to use instruments such as the flat magnetic field 180° pair spectrometer (§ 5.3) which employ converters of much larger area. The 180° pair spectrometer, however, is not restricted to this application and has, in fact, also been used to measure γ-rays emitted from the targets of charged particle accelerators.

Before proceeding to a discussion of the properties of various pair spectrometers, we shall consider, in general terms, the relative merits of the external and internal pair conversion process for γ-ray detection. This question has been discussed for lens-type spectrometers by Bame and Baggett (1951) and by Bent, Bonner and Sippel (1955). We shall assume that the same γ-ray source is used for both external and internal pair production. For simplicity the source will be assumed to be sufficiently thin that the scattering and energy loss of internal pair electrons in it is negligible, a situation which usually obtains if the source is the target of an accelerator. The counting efficiencies ϵ in counts per quantum at the maximum of the line may be written :

for external pair conversion : $\epsilon_{ex} = \Omega_s t \beta_{ex} B,$

and for internal pair conversion : $\epsilon_{int} = \alpha_{int} A,$

where Ω_s is the fraction of the sphere subtended at the source by the external pair converter, t is the converter thickness in atoms per cm^3, β_{ex} is the pair conversion cross section in cm^2 per atom, α_{int} is the internal pair conversion coefficient in pairs per quantum, and A and B are the fractions of the pairs produced in internal and external conversion which possess the momenta and directions required to allow their transmission through the momentum analyser at the magnetic field corresponding to the maximum counting rate. In order to compare these efficiencies, we make two simplifying assumptions : (*a*) that the distribution of γ-ray directions and of the total momentum vectors of the internal pairs from the source are isotropic and (*b*) that for external conversion the converter is placed very near to or in contact with the γ-ray source so that the solid angle subtended by the converter at any point on the source is 2π steradians. Then, provided the angular distributions of the

positron and electron in the two processes are not too dissimilar, we have $\Omega_e B \simeq A$. Thus, ϵ_{ex} will equal ϵ_{int} when $t\beta_{ex}$ is comparable with α_{int}. For a typical γ-ray energy of 5 MeV and for a lead radiator β_{ex} is $7{\cdot}5 \times 10^{-24}$ cm^2/atom (see for example Davisson and Evans 1952). The internal conversion coefficient for 5 MeV E1 γ-rays is approximately 2×10^{-3} per quantum (Rose 1949). Hence, we see that the efficiencies will be of the same order of magnitude if the radiator thickness is about $2{\cdot}7 \times 10^{20}$ atom/cm^2. However, a radiator of this thickness may seriously affect the resolution since the mean energy loss for the electron pair is about 90 kev and the mean scattering angle for a 2 MeV electron produced in the centre of the radiator is about 0·6 radian. The decision to use internal or external pair formation in practice therefore depends on a discrimination between high counting rate and poor resolution obtainable with external conversion, and good resolution and low counting rate obtainable with internal conversion. It was found by Bent, Bonner and Sippel (1955) in studying the γ-rays from light nuclei that for values of resolution poorer than 5%, external conversion gave the highest counting rate while for values of resolution better than 5%, internal conversion was most suitable. This rule is no longer valid if the bombarding particles are neutrons rather than charged particles because targets which are thick for internal conversion electrons may still be thin for neutrons. Under these conditions the greater efficiency at a given resolution may be achieved with external conversion.

5.2. *Lens-type Pair Spectrometers*

In pair spectrometers of lens type the positrons follow orbits through the lens which spiral in a direction opposite to that of the electrons. A pair spectrometer in which the positrons only were detected was used by Thomas and Lauritsen (1952) in the study of various light element nuclear reactions. A vast improvement in resolution and signal to background ratio is achieved if both the electron and the positron are detected in coincidence in the same spectrometer. If a given pair originates on the axis of the instrument and if the momentum division is nearly equal, the two trajectories will arrive at a common focus whose area will depend on the optics of the lens and the resolution conditions under which the spectrometer is operated. Two methods are commonly used to effect the separate detection of positive and negative electrons at the end of their trajectories. These are off-axis separation and statistical separation.

5.2.1. *Off-axis separation.*

The off-axis method was first described by Siegbahn and Johansson (1950). The instrument consists of a lens of such design that an image rotation of 90° is achieved. By placing the source at a position removed from the axis, well separated images are obtained for positrons and electrons of a given momentum. For pair events in which nearly equal momentum division occurs for the two particles, the images are in a line perpendicular to the plane containing the source and the axis of the spectrometer and are equidistant from the axis. This instrument has the advantage that a complete separation of the two charges is obtained. However, since this is achieved by using a point focus the transmission may be less than that for an instrument designed to produce a ring focus (see § 5.2.2) or an intermediate image focus (Slätis and Siegbahn 1949).

5.2.2. *Statistical separation.*

For the statistical separation method, which was discovered independently by Bame and Baggett (1951) and by Alburger (1952), the lens spectrometer is adjusted to give a ring focus or an intermediate image focus. The detector is placed beyond the ring baffle at a point where the converging trajectories of positrons and electrons produce a circle of confusion of small diameter. The detector consists of two semicircular scintillators separated by a lead or tungsten screen about $\frac{1}{8}$ in. thick. The two scintillators are biased to exclude pulses from annihilation radiation and their outputs are fed into a fast coincidence circuit. For transmitted electron pairs

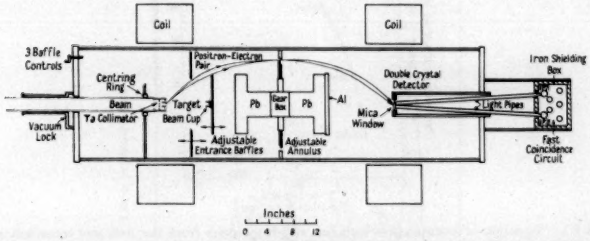


Figure 5.1. Schematic diagram of the intermediate image statistical separation pair spectrometer (Alburger 1956).

of nearly equal momentum division the positron and electron will strike different segments of the split detector in statistically one-half of the cases and the same segment in the other half.

The best performance for a statistical separation spectrometer appears to have been achieved by Alburger (1956) who used an instrument patterned after the intermediate image focusing β-spectrometer of Daniel and Bothe (1954). A diagram of the instrument is given in figure 5.1. The detectors were plastic scintillators. The associated electronic circuits included pulse height discriminators which were used to reject pulses from the detectors smaller than a chosen limit and a fast coincidence circuit with a resolving time of about 10^{-9} sec. A similar instrument has been developed at Rice Institute (Bent *et al.* 1955 a, b, Ranken *et al.* 1958). With an entrance annulus subtending at the source an angle of 4·4° centred at 51° to the spectrometer axis, the transmission of Alburger's spectrometer was 3% of 4π and the line width for conversion electrons was 1%. In an experiment using the 6·06 MeV nuclear pair line from ^{16}O it was found that between the line widths 0·56% and 2·5% the efficiency changed from $0{\cdot}07 \times 10^{-5}$ to $6{\cdot}3 \times 10^{-5}$ pair counts per transition in the manner shown in figure 5.2. From published results of measurements made with internal pairs (Bent *et al.* 1955 b) it may be deduced that the efficiency of the Rice Institute spectrometer was two or three times lower than Alburger's for 2·5% resolution at 6 MeV. The variation of the peak counting efficiency as a function of energy for the Rice instrument using internal pairs was calculated by Bent *et al.* (1955 b) from the differential cross section formulae of

Rose (1949). The resulting functions for various multipole orders and for 2·5% energy resolution are shown in figure 5.3. The relative efficiency varies by a factor of five from the minimum useful energy, about 2·7 MeV, to 12 MeV. A similar energy dependence applies to Alburger's spectrometer.

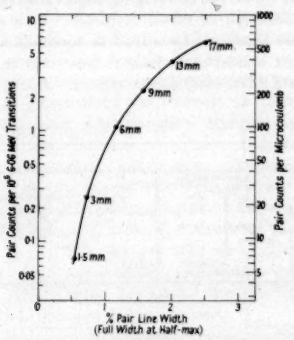


Figure 5.2. Variation of transmission with line width for pairs from the 6·06 MeV transition in ^{18}O produced by bombardment of a thick CaF_2 target with 2·2 MeV protons (Alburger 1956). The beam spot was 1 mm in diameter. The numbers on the curve give the widths of the adjustable annulus of figure 5.1.

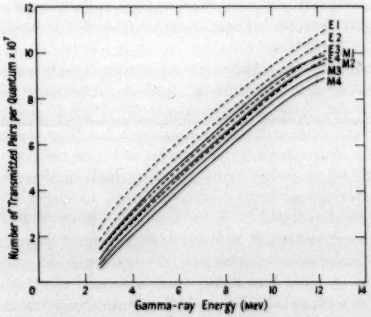


Figure 5.3. Theoretical energy dependence of the transmission of the lens spectrometer for various multipoles in internal pair conversion (Bent, Bonner, McCrary, Ranken and Sippel 1955).

The optimum useful resolution, with source strengths available, for Alburger's instrument appears to be about 1% although as shown by figure 5.2 resolutions as high as 0·56% are obtainable. The best resolution employed with the Rice Institute spectrometer in studying γ-rays from charged particle reactions was 1·8%.

An appreciation of the line shape and of the intensity of the background observed with the lens pair spectrometer is obtained from examination of figure 5.4 which shows a portion of the γ-ray spectrum from the $^{10}B+d$ reaction observed by Bent and co-workers at 1·8% resolution. Alburger, using the 6·06 MeV nuclear pair line from ^{16}O with pulse height discriminators set to include essentially the whole peak of the electron detector pulse height response curve, achieved a signal to background ratio of 80. The background was attributed mostly to scattered electrons which managed to register in both detectors. With the discriminator setting increased so that many more low energy electron pulses were rejected, the

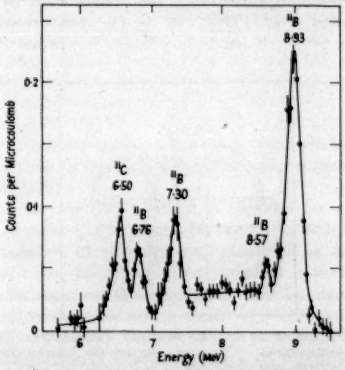


Figure 5.4. Internal pair spectrum from the bombardment of a thick ^{11}B target with 2·0 MeV deuterons. Resolution 1·8% (Bent, Bonner, McCrary, Ranken and Sippel 1955).

overall efficiency was reduced to one third while the background, then only a few times the random coincidence rate, was reduced to about one six-hundredth of the peak intensity.

The lens pair spectrometer does not appear to have been used extensively as an absolute instrument for measurement of γ-ray intensities or energies. The absolute efficiency of the Rice Institute instrument has been obtained theoretically as shown in figure 5.3, but the efficiency calibration adopted for the instrument is that obtained by measurement of γ-rays of known absolute intensity (Bent *et al.* 1955 b). The efficiency calibration of Alburger's instrument was obtained by counting all electrons (both positive and negative) transmitted from the ^{16}O 6·06 MeV nuclear pair line under spectrometer conditions for which the absolute transmission for conversion electrons had been previously determined using the 1·0639 MeV conversion line from a calibrated source of ^{207}Bi. Bent and co-workers calibrated the energy scale of their instrument using the ^{16}O pair line and thence quoted γ-ray energies with a minimum probable error of ± 20 keV (Ranken *et al.* 1958). Alburger measured the energy of the ^{16}O pair line itself relative to the 1·0639 MeV K-conversion line from ^{207}Bi and obtained $6·065 \pm 0·009$ MeV.

5.3. *Flat Field* 180° *Pair Spectrometer*

5.3.1. *Theory.*

The earliest experiments with a flat field magnetic spectrometer involved positron detection only (Alichanov and Dzhelepov 1938, Alichanov and Latyshev 1938), and the first measurements using the method of electron positron coincidences were made by McDaniel, von Dardel and Walker (1947) and Walker and McDaniel (1948). Spectrometers similar to theirs have been employed in the measurement of high energy γ-rays following charged particle reactions (Terrell 1950, Terrell and Phillips 1951, Rae, Rutherglen and Smith 1951) and in the measurement of neutron capture γ-rays (see compilation by Bartholomew and Higgs 1958, also Balzer *et al.* 1958). In the (n, γ) experiments, because of the high intensity of the γ-rays, it has been possible to operate the instrument with

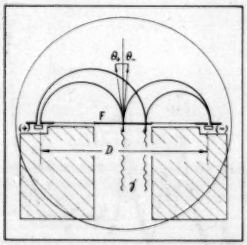


Figure 5.5. Schematic diagram of the 180° flat field pair spectrometer. The circle represents the pole face of the electromagnet. The γ-rays entering the vacuum chamber from the bottom of the figure pass between lead shields and strike the converter foil F. Electron pairs emitted at small angles θ to the γ-ray direction are bent round by the field and enter the detectors through apertures separated by the distance D.

an energy resolution to 1% or better and its performance in this application has been extensively studied (Kinsey and Bartholomew 1953, Bartholomew, Campion and Robinson 1960).

The operation of a flat field 180° pair spectrometer depends on the nearly linear relationship between the momentum and total energy of a particle moving in the region of relativistic velocities. Under this condition the sum of the moduli of the momenta $p_+ + p_-$ (hereafter called the total momentum) of an electron pair of total energy $\alpha = E_\gamma/m_0c^2$ is almost independent of the division of momentum between the two particles and a very nearly linear relation exists between the total momentum and the γ-ray energy. To a sufficient approximation the γ-ray energy E_γ is equal to the total electron energy. The nuclear recoil energy for pairs produced in the field of a heavy nucleus, e.g. gold, is of the order of an electron volt and may be neglected. In the geometry of figure 5.5, pairs produced by γ-ray quanta of energy α from different parts of the broad radiator F will have nearly equal values of $H(\rho_+ + \rho_-)$ where H is the field strength and ρ_+ and ρ_- are the radii of curvature of the positron and electron trajectories respectively. Thus for pairs ejected within a small range of polar angles $\theta_\pm \simeq 0$, an increase in

the coincidence counting rate between the two detectors will be observed when the field is such that $D < 2(\rho_+ + \rho_-) < D + 2a$, where D is the distance between the inner edges of the slits covering the detectors and a is the slit aperture width. In order for a coincidence to be detected it is also necessary for the electrons to satisfy various purely geometrical requirements imposed by the detector apertures, the principal condition being :

$$\frac{D - Y_+}{Y_+ + a} \leqslant \frac{p_- \cos\theta_-}{p_+ \cos\theta_+} \leqslant \frac{D - Y_+ + a}{Y_+} \qquad \ldots\ldots(5.1)$$

where Y_+ is the distance from the point of origin to the inner edge of the positron slit.

It may be shown (Kinsey and Bartholomew 1953) that, in the limit, pairs produced by a monoenergetic γ-ray with $\theta_+ = \theta_- = 0$ and $p_+ = p_-$ will have the greatest total momentum and, if ejected from the midpoint of the radiator, will be detected at H_0, the maximum possible value of the magnetic field strength for which coincidences from that γ-ray can occur. Pairs with $p_+ \neq p_-$ or with $\theta_\pm \neq 0$ are detectable only at field strengths less than H_0. The contributions of all pair events produced by a monoenergetic γ-ray beam and satisfying the geometrical requirements form a peak or line in the counting rate spectrum (counting rate as a function of field strength).

A calculation of the line shape, which is essential for a precise knowledge of the relationships between the transmission, resolution and γ-ray energy and for the absolute determination of energies and intensities, is difficult to carry out in an exact form including the effect of an extended radiator. However, some insight into the performance of the instrument may be obtained from the consideration of idealized geometrical arrangements in which the radiator is confined to a point symmetrically placed between the detectors. Two such geometries involving quite different types of detector aperture are discussed below. For both geometries the theoretical line shape can be calculated with precision. One of these apertures represents an idealization of the type of aperture which would appear to give the highest transmission for low resolution operation of the instrument while the other does the same for very high resolution.

Other factors affecting the line shape and resolution of the spectrometer are scattering and energy loss of electrons in the radiator, inhomogeneities in the magnetic field and scattering from the walls of the vacuum chamber and the jaws of the slits. For further consideration of such effects the reader is referred to the original publications (Kinsey and Bartholomew 1953, Bartholomew *et al.* 1960).

5.3.1.1. *Idealized geometries.*

In the first, low resolution (thin slit) arrangement the apertures are infinitely long and are formed by parallel jaws of zero thickness (Kinsey and Bartholomew 1953). The apertures are parallel to the magnetic field direction and the direction of incidence of the γ-radiation is normal to the plane containing the apertures (figure 5.6). The coordinate y_+ of the intersection with this plane of the trajectory of a positron defined by the parameters (p_+, θ_+, ϕ_+) is given by

$$y_+ = \frac{2p_+ c}{eH} \cos\theta_+ \qquad \ldots\ldots(5.2)$$

where e is the electronic charge in e.s.u. and c is the speed of light. It is clear from this expression and the geometry of figure 5.6 that there exists a field H such that any positron trajectory can enter the positive aperture regardless of ϕ_+, and that the only restriction imposed by these apertures on the detection of a positron–electron pair (p_+, θ_+, ϕ_+ ; p_-, θ_-, ϕ_-) at some value of $H \leqslant H_0$ is given by (5.1). It follows that the line produced by these slits has a tail extending indefinitely below the field H_0, the counting rate in the tail depending on the relative numbers of electron pairs with different combinations of the parameters $p_\pm$, and $\theta_\pm$, i.e. on the differential cross section for pair production. In practice the distribution in θ_+ and θ_- is also dependent on the degree of collimation of the γ-rays and their angle of incidence with respect to the plane containing the detectors.

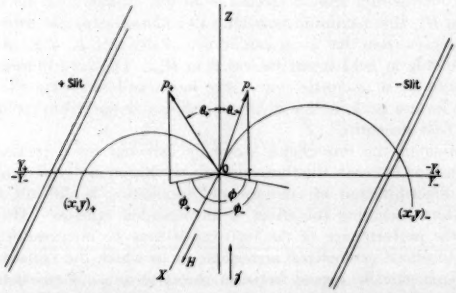


Figure 5.6. Idealized geometry for thin slits for the 180° flat field pair spectrometer.

In the second, high resolution (curved slit) arrangement (Bartholomew *et al.* 1959) the apertures are formed by thin straight-edge jaws on the side nearest the radiator and doubly curved thick jaws on the outer side, as shown in figure 5.7. The outer jaws have a radius of curvature $R = d/2$ in the plane perpendicular to the magnetic field and a radius of curvature $R = \pi^2 d/4$ in the plane perpendicular to the γ-ray direction, where d is the distance from the point of origin O to the inner edge of the slit. These apertures restrict the angles θ_+ and θ_- to a maximum value given by $\cos\theta_{\max} = 1-\delta$, where δ is the ratio of the aperture width in the plane of symmetry to d. Five limiting positron orbits (one with $\theta_+ = 0$ and four with $\theta_+ = \theta_{\max}$) are shown in figure 5.7.

For the purpose of describing the line shape we introduce the following parameters : The field parameter $\omega = (H_0-H)/H_0$ measures the displacement of the magnetic field H from H_0, the maximum value at which coincidences from a given γ-ray can occur. The momentum parameter $z = |p_0-p_+|/p_0$ measures the displacement of the momentum of either member of the pair from p_0, the momentum at equal momentum division. To a close approximation, near equal momentum division, we have $p_+ + p_- = 2p_0$. The effects of large departures from p_0 are considered below.

The line profile of a γ-ray of energy α produced by the ' thin ' slits of figure 5.6 is given (Kinsey and Bartholomew 1953) by

$$I_1(\alpha, \omega, \delta) = 2\int_0^{\omega}\int_{1-\omega+z}^{y_1}\int_{1-\omega-z}^{x_1} G(x, y, \alpha)\, dx\, dy\, dz \qquad \ldots\ldots(5.3)$$

in which we have neglected quantities of second order in ω and z and where $x = \cos\theta_+$, $y = \cos\theta_-$, $x_1 =$ lesser of $1-\omega-z+\delta$ and 1, and $y_1 =$ lesser of $1-\omega+z+\delta$ and 1. The line profile produced by the curved slits of figure 5.7 is given (Bartholomew *et al.* 1960) by

$$I_2(\alpha, \omega, \delta) = 2\int_0^{\zeta}\int_{1-\omega+z}^{1}\int_{1-\omega-z}^{1} G(x, y, \alpha)\, dx\, dy\, dz \qquad \ldots\ldots(5.4)$$

where $\zeta =$ lesser of ω and $\delta-\omega$. The function $G(x, y, \alpha)$ is the differential cross section for pair production. It is important to note that both types of idealized aperture impose no restrictions on the azimuthal angles ϕ_+ and ϕ_-. Therefore for

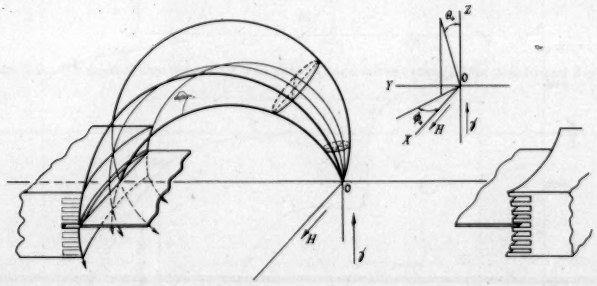


Figure 5.7. Aperture geometry for curved slits. The apertures transmit only pair electrons emitted from O within a cone determined by the slit parameter δ. In practice the aperture size is adjusted by sliding the movable outer jaw over the fixed inner jaw. The channels in the outer jaw (shown in detail in the right-hand side only) reduce electron scattering.

the cross section $G(x, y, \alpha)$ it is convenient to use a form in which integration over these angles has already been performed. The function $G(x, y, \alpha)$ is strictly also a function of the momentum division z. However, for the present geometry (central point radiator) only small departures from equal momentum division can occur and it is assumed that over this interval $G(x, y, \alpha)$ is, to a close enough approximation, independent of z. By integrating the expression for the differential cross section for pair production given by Bethe and Heitler (1934) over the azimuthal angle ϕ we derive, for the case of equal momentum division,

$$G(x, y, \alpha) = -\frac{Z^2 r_0^2}{8\alpha^2 S^2 137} P(x, y, \alpha) \qquad \ldots\ldots(5.5)$$

where

$$P(x,y,\alpha) = \frac{4}{(S-x)(S-y)} - \frac{1}{(A^2-B^2)^{1/2}}\left[\frac{1-x^2}{(S-x)^2}+\frac{1-y^2}{(S-y)^2}+\frac{4(S^2+A)}{(S-x)(S-y)}\right]$$
$$+\frac{S^2A}{(A^2-B^2)^{3/2}}\left[\frac{1-x^2}{(S-x)^2}+\frac{1-y^2}{(S-y)^2}-\frac{2(2-x^2-y^2-2A)}{(S-x)(S-y)}\right]$$
$$A = S^2+\tfrac{1}{2}+\tfrac{1}{2}xy-S(x+y)$$
$$B = \tfrac{1}{2}[(1-x^2)(1-y^2)]^{1/2}$$
$$S = [1-(2/\alpha)^2]^{-1/2}$$

and where Z is the atomic number of the radiator material and r_c is the classical radius of the electron.

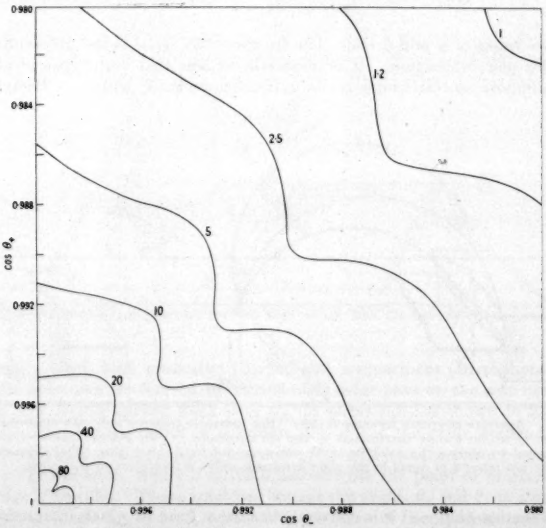


Figure 5.8. A plot of the distribution $P(x, y, \alpha)$, eqn (5.5) for $\alpha = 16$. The amplitude is zero at $\cos\theta_+ = \cos\theta_- = 1$ and passes through a saddle point (not shown) on the diagonal between $\cos\theta_+ = 1$ and the 80 contour.

A plot of $P(x, y, \alpha)$ for $\alpha = 16$ is given in figure 5.8. The line profiles $I_1(\alpha, \omega, \delta)$ and $I_2(\alpha, \omega, \delta)$ for $\alpha = 16$ and $\delta = 0{\cdot}010$ are compared in figure 5·9. The high field sides of the lines are identical, an effect which arises because the apertures have identical inner jaws. However, while the $I_1(\alpha, \omega, \delta)$ profile has a tail which extends indefinitely toward lower values of H, the line $I_2(\alpha, \omega, \delta)$ cuts off abruptly at $\omega = \delta$. It is clear that the transmission–resolution ratio given by the curved slits is quite superior to that given by thin slits ; the maximum of the $I_2(\alpha, \omega, \delta)$

line for 1·4% resolution would be some four times greater than that for $I_1(\alpha, \omega, \delta)$. It may be shown by simple graphical arguments (Bartholomew *et al.* 1960) that both the high-field and low-field sides of the line obtained with the curved slits have

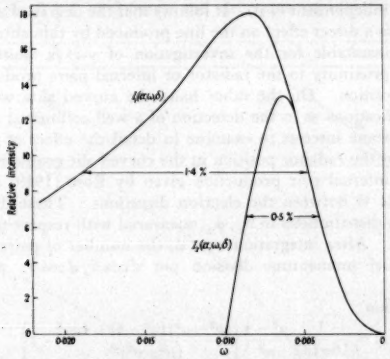


Figure 5.9. Theoretical line shapes for thin slits (I_1) and curved slits (I_2) for $\alpha = 16$ and $\delta = 0{\cdot}010$.

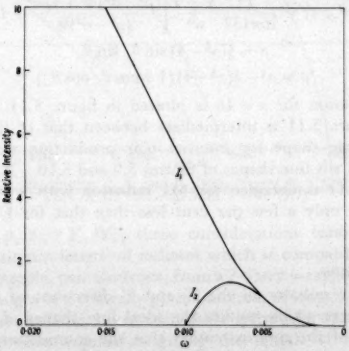


Figure 5.10. Theoretical line shapes for a flat differential cross section for pair production for thin slits (I_1) and for curved slits (I_2) for $\delta = 0{\cdot}010$.

the steepest possible slopes and therefore these slits must give the optimum transmission–resolution ratio obtainable with a point radiator.

In order to separate the effects of the differential cross section $G(x, y, \alpha)$ and the geometry of the apertures in determining the line shape for the two types of aperture we shall now consider the line shapes for a flat distribution ($G(x, y, \alpha)$ equal to unity). The line profiles obtained for $\delta = 0{\cdot}010$ are shown in figure 5.10.

That for the thin slits is found to increase with decreasing H without forming a maximum while that for the curved slits is very similar to the line for the real differential cross section. These results demonstrate that the line profile for thin slits depends strongly on the electron angular distribution while that for curved slits is essentially independent of it. It follows that the degree of collimation of the incident γ-rays has a direct effect on the line produced by thin slits. Such slits are therefore quite unsuitable for the investigation of γ-rays emitted by a source situated in close proximity to the radiator or internal pairs produced by a source at the radiator position. On the other hand the curved slits would function as well in these applications as in the detection of a well collimated γ-ray beam.

It may be of some interest to examine in detail the effect of placing a source of internal pairs at the radiator position in the curved slit geometry. The angular distributions for internal pair production given by Rose (1949) are expressed in terms of the angle Θ between the electron directions. These distributions may be transformed to distributions in $\theta_\pm$, $\phi_\pm$, measured with respect to the space fixed axes of figure 5.7. After integration over ϕ, the number of pairs per unit energy interval near equal momentum division per $d\cos\theta_+ d\cos\theta_-$ per quantum of energy α is :

for M1 radiation

$$G_{M1}(\theta_+, \theta_-, \alpha) = \frac{1}{16\pi 137}\frac{\alpha^2-4}{\alpha^3}\left[\frac{(\rho^2-\sigma^2)(\alpha^2-4)+4\rho\alpha^2}{(\rho^2-\sigma^2)^{3/2}}-1\right] \qquad \ldots\ldots(5.6)$$

and for E1 radiation

$$G_{E1}(\theta_+, \theta_-, \alpha) = \frac{1}{16\pi 137}\frac{\alpha^2-4}{\alpha^3}\left[\frac{(\rho^2-\sigma^2)\alpha^2+4\rho\alpha^2}{(\rho^2-\sigma^2)^{3/2}}+1\right] \qquad \ldots\ldots(5.7)$$

where $$\sigma = \tfrac{1}{2}(\alpha^2-4)\sin\theta_+ \sin\theta_-$$

and $$\rho = \alpha^2 - \tfrac{1}{2}(\alpha^2-4)(1+\cos\theta_+\cos\theta_-).$$

The E1 distribution for $\alpha = 16$ is plotted in figure 5.11. Since the angular dependence of figure 5.11 is intermediate between that of figure 5.8 and a flat distribution, the line shape for internal pair production would lie somewhere between the curved slit line shapes of figures 5.9 and 5.10. A distribution similar to that in figure 5.11 is obtained for M1 radiation with $\alpha = 16$, the differential cross section being only a few per cent less than that for E1 radiation, for the angle range of interest.

5.3.1.2. *Finite radiator.*

Extension of the radiator in the X and Y directions of figures 5.6 and 5.7 introduces two effects which modify the ideal line shapes of figure 5.9. These are (i) breakdown of the approximation that the momentum of the electron is proportional to its energy, (ii) failure of the geometric conditions which strictly apply only to the point O.

(i) It can easily be shown that the departure from linearity of the momentum–energy relationship does not contribute in an important way to the line width except under the combined conditions of high resolution ($<1\%$) and low γ-ray energy ($E_\gamma < 3$ MeV). The relation between the total momentum of the positron and electron and the γ-ray energy is contained in the equation

$$c(p_+ + p_-) = (E_+^2 - m_0^2c^4)^{1/2} + (E_-^2 - m_0^2c^4)^{1/2} \qquad \ldots\ldots(5.8)$$

where E_+ and E_- are the electron energies ($E_+ + E_- = E_\gamma$) and all other symbols have their usual meaning. The diminution of total momentum as a function of the departure from equal momentum division is given in figure 5.12 for three values of γ-ray energy. For a momentum asymmetry of 30% the decrease of total momentum is only 0·6% at 2·6 MeV and decreases rapidly for higher energies.

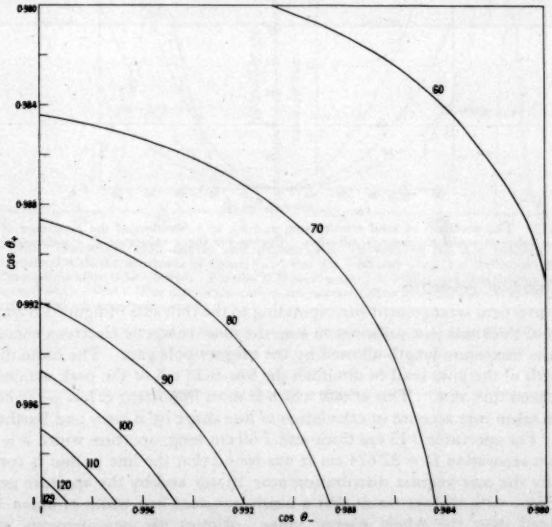


Figure 5.11. Plot of the distribution in θ_+ and θ_- (with respect to a space fixed axis), of internal pairs of E1 radiation with $\alpha = 16$ (5.7).

Since, from (5.1) $p_+/p_- \simeq Y_+/Y_-$, these considerations lead immediately to an estimate of the maximum length of radiator which is compatible with any desired resolution. For example, pair electrons, from a 2·6 MeV γ-ray, ejected in the forward direction from the edges of a centrally placed radiator of length $D/3$ will be recorded at a field strength about 0·6% lower than those ejected forward from the centre of the radiator. A precise expression giving the length of radiator effective in contributing to the counting rate at the maximum of the line is given by Kinsey and Bartholomew (1953).

(ii) The line profiles corresponding to points of origin removed from the centre of symmetry are difficult to calculate and no analytical description will be attempted here. However, it is apparent that the profile produced by the curved slits will be more sensitive than that produced by thin slits to translations of the point of origin. This is particularly true for translations of the point of origin along the X axis to which the thin slits (being infinitely long) are completely insensitive. It

has been shown experimentally (Kinsey and Bartholomew 1953) that for thin slits of finite length the line profile is quite insensitive to the point of origin of the pairs. On the other hand experiments have shown that with curved slits the line profile is quite sensitive to the size of the radiator (Bartholomew *et al.* 1960).

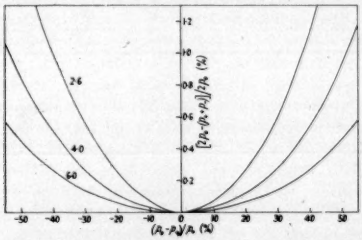


Figure 5.12. The variation of total momentum, $p_+ + p_-$, as a function of the departure of p_+ from p_{+0}, the value at equal momentum division. Curves for three values of E_γ (MeV) are shown.

5.3.2. *Practical geometries.*

The practical arrangement corresponding to the thin slits of figure 5.6 consists of slit jaws of thickness just sufficient to stop the most energetic electrons encountered and of the maximum length allowed by the magnet pole gap. The finite thickness and length of the jaws tend to diminish the low-field tail of the peak obtained with the idealized thin slits. This effect, which is most important at low γ-ray energies, has been taken into account in calculation of line shape by Kinsey and Bartholomew (1953). For apertures 0·15 cm thick and 7·60 cm long, aperture width $a = 0{\cdot}2$ cm and mean separation $D = 22{\cdot}674$ cm it was found that the line profile is controlled largely by the pair angular distribution near 10 MeV and by the aperture geometry near 2·7 MeV with the net result that a nearly constant line width of about 120 keV is obtained over the whole energy range. Almost all measurements of γ-ray spectra with the 180° magnetic pair spectrometer have been made with this type of aperture or with an arrangement in which the sensitive areas of the electron detectors themselves acted as the defining apertures. As mentioned above, this geometry is best suited for applications where high transmission is needed and low resolution can be tolerated. It is poorly suited for detection of a divergent γ-ray beam. A modification described by Alburger (1955) involves curving the radiator in such a way as to compensate for the effects of a divergent beam. As yet, this arrangement does not appear to have been tried in practice.

The curved slits of figure 5.7 may be used to advantage wherever very high resolutions are required. It has been shown experimentally (Bartholomew *et al.* 1960) that even with radiators of considerable area these slits give superior performance to other types at very high resolution. In figure 5.13 are shown both the theoretical and experimental line profiles for two γ-rays measured with the apertures of figure 5.7. For both measurements the radiator is of such dimensions (given in the caption) that it could hardly be considered a very close approximation to a point. However, at 0·9% resolution, the line shape is not seriously distorted

and the low-field tail is not excessively large, and for both curves the shape of the high field side of the peak is well preserved.

For intermediate resolutions, a third type of slit (thick slit) has been found to give a transmission greater than either of the other types. Both the inner and

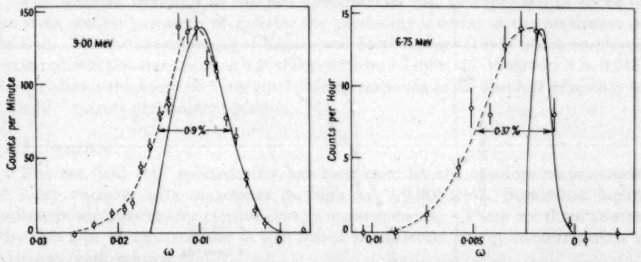


Figure 5.13. Theoretical (full curve) and experimental (broken curve) line profiles for γ-rays measured with the apertures of figure 5.7. For the 9·00 MeV γ-ray $\delta = 0{\cdot}016$ and the radiator dimensions were $0{\cdot}5d \times 0{\cdot}44d$. For the 6·75 MeV γ-ray $\delta = 0{\cdot}0024$ and the radiator dimensions were $0{\cdot}44d \times 0{\cdot}09d$. The theoretical curves are normalized to have the same maximum counting rate as the experimental curves.

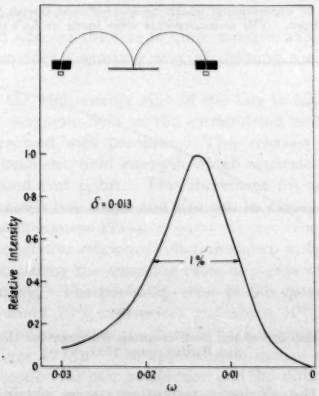


Figure 5.14. Schematic diagram of the aperture geometry and a typical line profile at 6·75 MeV for 'thick' slits.

outer jaws of this aperture are thick (approximately equal to five times the maximum gap width) and the aperture is curved with a radius equal to d in the plane perpendicular to the magnetic field direction (see inset, figure 5.14). The jaws have no curvature in the direction of the magnetic field. The line profile calculation for these apertures is very complicated. However, it is found empirically (Bartholomew

et al. 1960) that the line has a symmetrical shape and a short tail on the low-field side as shown in figure 5.14. The sensitivity of these apertures to both radiator area and degree of collimation of the γ-ray beam is intermediate to that of the other two types.

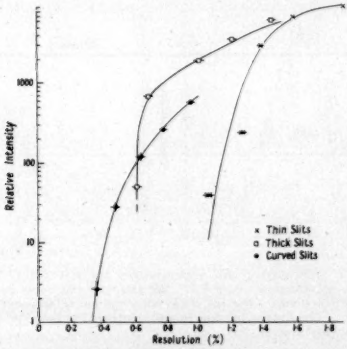


Figure 5.15. Comparison of experimental transmission–resolution curves for three types of detector aperture. The measurements were made at 6·75 MeV.

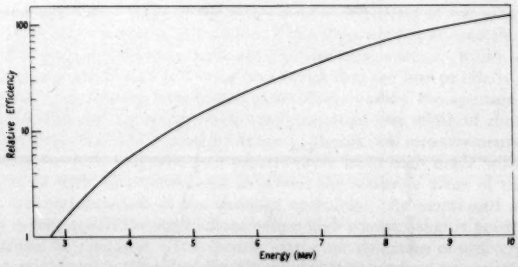


Figure 5.16. Energy dependence of the peak counting efficiency for thin slits $\delta = 0{\cdot}018$ (Kinsey and Bartholomew 1953).

A comparison of the efficiency–resolution curves obtained experimentally at 6·75 MeV for the three types of slit is shown in figure 5.15. The details of the experimental arrangement for these measurements are given by Bartholomew *et al.* (1960). The curves show the ranges of resolution where each of the slits gives superior transmission.

The energy dependence of the efficiency for the thin slits of finite length described above is given in figure 5.16. The variation with energy is much more rapid than that of the total pair cross section itself. Such a result is expected

since at low energies the average acceptance angle θ determined by the finite thickness and length of the apertures is less than the mean angle of emission of the pairs. A similar energy dependence of the transmission is found for the apertures of figures 5.7 and 5.14.

The absolute efficiency of the pair spectrometer can be expressed in terms of the cross section per atom of radiator for producing a count at the maximum of the line. For the spectrometer of Kinsey and Bartholomew (1953) which employed a gold radiator this transmission was observed to be 7·7 mbn at 7·38 MeV for $\delta = 0{\cdot}018$. For a radiator thickness of 5 mg cm^{-2} this corresponds to an absolute efficiency of $1{\cdot}1 \times 10^{-7}$ counts per incident photon.

5.3.3. *Precision.*

The flat field 180° spectrometer has been used for the absolute measurement of γ-ray energies with accuracies as high as $\pm 0{\cdot}006$ MeV. Somewhat higher accuracies are possible for relative energy measurements. There are three reasons why this type of spectrometer is well suited for absolute energy measurements of relatively high accuracy:

(*a*) The energy determination involves a precise knowledge of only two quantities: the distance between the inner edges of the apertures and the magnetic field at a representative position on the line. The magnetic field is flat and therefore easily measured. Small irregularities in the field strength do not seriously alter the line shape particularly on the high-field edge.

(*b*) The high-field edge of the line is quite straight and when extrapolated to zero counting rate provides a sensitive way of defining the magnetic field at the line.

(*c*) The shape of the high-energy side of the line is easily calculated and the relation between the magnetic field at the extrapolated end point and the γ-ray energy may be determined with precision. This relation involves a knowledge of the increment of magnetic field strength which separates the field H_0 and the field at the extrapolated end point. This increment (or toe correction) can be calculated from the known line shape and also can be obtained by direct measurement (Kinsey and Bartholomew 1953).

The validity of the relative efficiency function shown in figure 5.16 was checked experimentally by comparing the counting rates of γ-rays of different energy and known absolute intensity. The probable error of the curve relative to 2·75 MeV was estimated to be about 20% near 5 MeV and about 10% near 7 MeV while for higher energies the error may be as much as 30% (Kinsey and Bartholomew 1953, Bartholomew and Higgs 1958). The approximate nature of the expressions for the differential cross section for pair production and the difficulty of calculating the efficiency at the maximum of the line for slits of practical design tend to make the flat field 180° pair spectrometer unattractive as an absolute instrument for measuring photon intensities.

5.4. *Comparison of Pair Spectrometers*

We shall now briefly consider the relative merits of the lens-type and 180° flat field type of pair spectrometers and indicate the applications for which each seems best suited.

5.4.1. *Acceptance solid angle and transmission.*

It is of interest to compare the acceptance solid angles of the lens pair spectrometer and the 180° spectrometer of figure 5.7 since they give some insight into the applications for which the two types of instrument may be used to advantage. At 2·5% resolution the lens spectrometer of Bent *et al.* (1955 a, b) has an acceptance solid angle (for nearly equal momentum division) of 8·7% of 4π between cones with vertex angles of 31° and 40°. Theoretically, to give the same resolution the 180° spectrometer of figure 5.7 would have a maximum acceptance solid angle of 2·5% of 4π centred on the normal to the plane containing the radiator and detector apertures. Therefore, for pairs emitted isotropically into the spectrometer volume from a source or radiator of small area (either from internal pair production or from external pair production with the target situated immediately behind the radiator), the lens spectrometer has the better transmission at this resolution.† For very high resolutions, 0·6%, where finite source size and aberrations of the lens may begin seriously to affect the performance of the lens-type spectrometer, it is conceivable that the 180° flat field spectrometer could provide the better transmission. Up to the present time the low γ-ray intensities available from accelerators in general precludes the use of either spectrometer at these resolutions.

For well collimated γ-rays emitted from broad sources the 180° instrument clearly gives the best performance because (*a*) a broad radiator can be used and (*b*) pairs ejected into the small cone about the Z direction (figure 5.6) are transmitted while in the lens spectrometer only pairs directed at an angle of 30° to 50° from the spectrometer axis are detected. Since the majority of pairs are ejected into a small cone about the γ-ray direction the 180° spectrometer has the higher transmission in this application. A further increase in transmission for the 180° spectrometer may be accomplished by using multiple detectors (Walker and McDaniel 1948, Terrell 1950, Rae *et al.* 1950, Balzer *et al.* 1958).

5.4.2. *The efficiency function.*

The efficiency function for a lens spectrometer used for internal pair detection is given in figure 5.3 and that for a 180° spectrometer used for external pair detection in figure 5.16. It has already been mentioned that the rapid variation of the latter curve is due in large measure to the acceptance angle of the spectrometer which transmits the peak of the pair distribution at high energies and excludes it at low energies. On the other hand, the internal pair distribution with respect to the axis of the lens spectrometer is such that all directions in space are equivalent and the efficiency of the lens spectrometer has a variation with energy which resembles that of the internal pair conversion coefficient. If the 180° spectrometer were used to detect internal pairs, its counting efficiency would have a somewhat similar energy dependence. A slowly varying efficiency function is, of course, more desirable than one that varies rapidly since the corrections required to deduce relative intensities of γ-rays from the observed spectrum are smaller.

† *Note added in proof.* It is perhaps more realistic to compare an effective acceptance solid angle, $\Omega_1\,\Omega_2(\Theta)$ where Ω_1 is the usual geometrical acceptance solid angle for one of the pair particles and $\Omega_2(\Theta)$ is the acceptance solid angle for simultaneous transmission of the other particle at the mean pair emission angle Θ.

5.4.3. *Background.*

The probability of observing spurious coincidences caused by detection in one counter of scattered electrons or secondary γ-radiation produced in the other is much greater for the lens spectrometer than for the 180° spectrometer because the detectors are close together in the former and well separated in the latter. In addition, the probability of electrons and γ-rays being scattered into the acceptance solid angle is greater for the lens spectrometer. On the other hand, the 180° spectrometer is especially vulnerable to background caused by back-scattering of γ-rays after they have passed through the radiator, a problem which does not arise in the lens spectrometer. When used in an application where the target of an accelerator is placed at or near the radiator position, it is to be expected that the 180° spectrometer would have a troublesome background counting rate in each detector from neutrons generated in the target. This difficulty would be rather less serious in the case of the lens spectrometer since the electron detectors are situated at a somewhat greater distance from the target than they are in the 180° spectrometer.

§ 6. Summary

In this section we shall attempt to collect together some of the information which would assist the reader in deciding which spectrometer should be used for a particular type of measurement.

First let us summarize the data on the precision of the energy measurements which may be made with the different spectrometers ; at the same time the range of photon energies in which the instruments are useful will be indicated. Crystal-diffraction spectrometers are used to compare Bragg angles, and hence photon energies, with an accuracy to 3 parts in 10^5 to 3 parts in 10^4 for energies of 100 kev and 1 Mev respectively. In certain special cases lines have been measured with slightly higher precision ; these lines serve as energy standards for most of the other types of spectrometer. With conversion-electron spectrometers measurements of photon energies accurate to 1 part in 10^4 are fairly common in the range 50 kev to 3 or 4 Mev ; there is no inherent reason why work cannot be done at higher energies ; comparisons of electron momenta can be made to an accuracy of about 1 part in 10^5. With Compton spectrometers energy values accurate to 1 part in 10^4 have been obtained ; the useful energy range is approximately 0·5 Mev to greater than 10 Mev. Pair spectrometers have been used to make absolute measurements with an error of only about 6 kev in the energy range 3 Mev to 12 Mev ; slightly more accurate results are possible for relative energy measurements.

Although spectrometers which work with internally produced charged particles can be used to obtain γ-ray energies, only those which accept photons can be made to give direct measurements of γ-ray intensities. Measurements of the intensities of internally produced conversion electrons or pairs yield photon intensities only if the appropriate conversion coefficient is known for the nuclide in question. With a photon-accepting spectrometer γ-ray intensities can be obtained directly, once the instrument has been calibrated by the determination, either theoretically or experimentally, of its spectral efficiency function. The two-crystal diffraction spectrometer can be calibrated experimentally, by comparison

of the intensities in its one-crystal and two-crystal modes of operation; relative photon intensities have been measured with an accuracy to 5–10% and are found to agree with theory. A lens spectrometer with an external converter has been used by Thomas and Lauritsen (1952) to measure, with an accuracy to 25%, the relative intensities of γ-rays from nuclear reactions in light nuclei. For the simpler forms of the Compton spectrometer, e.g. the ritron, the spectral efficiency functions have been computed theoretically and tested experimentally; the discrepancy between theory and experiment in the region 1 to 3 MeV appears to be less than a few per cent. Relative intensity measurements have been made with the 180°, uniform-field, pair spectrometer with an accuracy to 10–30% in the range 3 to 8 MeV.

It often happens that more than one type of spectrometer can be expected to yield results of adequate accuracy in a particular energy range. In such circumstances one would like to choose the instrument which would allow the measurement to be made in the smallest interval of time. Unfortunately the time required to make a measurement depends upon a multitude of factors, so that it is very difficult to assign generally useful figures-of-merit to the different types of spectrometer. The presence of an appreciable background lengthens the time needed for a given measurement. Thus, although an instrument with a large intrinsic background may have a higher efficiency, it may still be inferior to one with a smaller true counting rate and a lower background. The true counting rate is proportional to the strength of the source and to the efficiency (the number of counts per photon at the peak of the line) of the spectrometer. Some spectrometers are designed to use sources of large volume (broad source instruments), while others are intended for use with sources of small volume (point- or line-source instruments); spectrometers of the latter class usually have compensatingly larger efficiencies. The choice of the type of spectrometer thus depends upon the amount of source material to be used. But an increase in source volume is usually associated with a worsening of the resolution. An increase in the source thickness —i.e. in the dimension parallel to the direction of emission of the photons or electrons—causes line broadening through energy degradation; while this is not a serious effect in photon-accepting spectrometers, it is most important in instruments using internally produced electrons. An increase in the source dimensions perpendicular to the direction of emission worsens the resolution by increasing spectrometer aberrations, in all cases except the flat-crystal diffraction spectrometers. In the same way attempts to improve the efficiency by increasing the acceptance aperture will result in a worsened resolution. Thus it is unprofitable to compare counting rates without at the same time taking resolution into consideration. Though the same precision of measurement with different spectrometers need not correspond to the same resolution, because the line-shapes are not all known with the same accuracy, we shall simplify the discussion by assuming that the different spectrometers are operated with roughly the same resolution; the comparison is complicated by the fact that the resolution often varies extremely rapidly with energy. Once the resolution has been fixed, the useful counting rate for a given type of spectrometer can be increased by increasing the size of the instrument; the counting rate usually varies as the square of the scale of the instrument. In comparing spectrometers of different types one must take into

account their relative sizes; because there is no unique criterion for choosing a fair ratio of sizes, the comparison of counting rates becomes even more arbitrary.

In spite of all the difficulties enumerated in the preceding paragraph, one can still make a useful, though very rough, comparison of spectrometer efficiencies if one groups together those that use the same type of source. First consider the instruments which use charged particles emitted from a source of small volume—i.e. the internal-conversion and the internal-pair spectrometers. Curves of efficiency against energy have been calculated by Bent *et al.* (1955 b; our figure 5.3) for a lens-type pair spectrometer at a resolution of 2·5%. We prefer to make a comparison at a better resolution, more suitable for precision spectroscopy. Hence

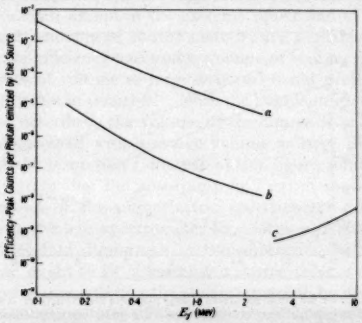


Figure 6.1. Efficiency curves for spectrometers using electrons or pairs emitted from a small source. Curves *a* and *b* refer to a lens spectrometer using internal conversion in an atom of atomic number 83 and 10 respectively, curve *c* to a lens-type internal-pair spectrometer. In each case the resolution is about 0·6%.

the Bent *et al.* curve for an E1 transition was renormalized, in accordance with Alburger's (1956) experimental data, to be consistent with a resolution of 0·56% at 6 MeV. The resulting curve is shown as *c* in figure 6.1. The internal pair coefficient is believed to be only weakly dependent on atomic number. A fair comparison can be made with a lens-type internal-conversion spectrometer, working at the same resolution. Though instruments capable of a somewhat higher transmission exist, we shall use the transmission–resolution curve of the Pasadena ring-focus spectrometer; at 0·56% resolution its transmission would be about 2·6%. Since the transmission is defined by the aperture of the spectrometer, there is a further reduction in intensity when one considers the exit-slit; for the Pasadena spectrometer we estimate that about 45% of the electrons entering the aperture pass through the exit-slit. The efficiency of an internal-conversion spectrometer is the product of the corrected transmission and the appropriate internal-conversion coefficient; the latter we take from the work of Rose *et al.* (1951) and Rose (1958). Curves *a* and *b* of figure 6.1 refer to E1 conversion in the K shells of atoms of atomic numbers 83 and 10 respectively. Because the two different types of spectrometer can use essentially the same source at a specified resolution, a

comparison of efficiency is equivalent to one of counting rate. We can conclude, therefore, that at a resolution of 0·56% the conversion-electron spectrometer is superior in counting rate to the internal-pair spectrometer even for an atomic number as low as 10, except possibly at the extreme upper end of our energy range. Because the efficiency–resolution curve of the former is flatter than that of the latter, the superiority of the conversion-electron type will be increased at better resolutions; at poorer resolutions the relative standing of the internal-pair instrument will be improved. The curves for the K-shell conversion of higher multipolarities lie above those for the E1 transition—often several orders of magnitude higher. Those for internal-pair production in transitions of higher multipolarity lie below the E1 curve. If the lens-type conversion-electron spectrometer were

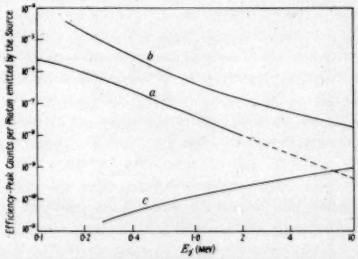


Figure 6.2. Efficiency curves for spectrometers using photons emitted from a small source. Curve *a* refers to a curved-crystal diffraction spectrometer, *b* to a lens spectrometer used with external conversion, and *c* to a Compton spectrometer.

replaced by a $\pi\sqrt{2}$ spectrometer, the efficiencies shown in curves *a* and *b* would be somewhat reduced, but there would be an overall improvement because of the increased source area (see § 3.4).

Another group of spectrometers which we shall consider consists of those instruments designed to work with photons emitted from a source of small volume. Because the different members of this group require sources of different shapes and sizes, fair comparison is difficult. Let us begin by comparing the efficiencies alone. (If the amount of source material available is very small, a comparison of efficiencies is all that is needed.) Curve *a* of figure 6.2 shows the energy dependence of a focusing quartz-crystal diffraction spectrometer (see figure 2.11 and § 2.4); the resolution is about 1·2% at 1 MeV. To represent the conversion-electron spectrometers we shall use a lens spectrometer with the converter arrangement of the example discussed in § 3.1. Assume a corrected transmission of 1·5%, corresponding to a resolution of 0·73% in the Pasadena ring-focus spectrometer if one ignores the energy broadening in the converter; the use of a uranium converter of 8 mg per cm² thickness should result in a resolution of the order of 1% at 1 MeV. The efficiency of such an external-conversion spectrometer is shown in curve *b* of figure 6.2. For the Compton spectrometers no curve of efficiency, as defined here, against energy appears to have been published, though

the efficiency of the ritron at 1 MEV is given as 4·5 coincidence counts per sec per curie. Curve *c* was obtained by renormalizing the theoretical curve from the design calculation of Lee-Whiting to the ritron efficiency at 1 MEV. Though the ritron resolution varies rapidly with energy, as shown in figure 4.6, its value at 2 MEV is about 3·5%. We shall not give a sample curve for a pair spectrometer in this group, since it is now generally agreed that one should use internal pairs under these conditions, except possibly at very poor resolutions. In the region of 1 MEV the diffracting spectrometer and the photoelectric spectrometer are roughly equal in efficiency and resolving power ; the resolution of the former deteriorates rapidly at high energies, that of the latter at low energies. In spite of its poorer resolution the Compton spectrometer is considerably inferior in efficiency to the photoelectric spectrometer, except at the extreme upper end of the energy range. However, when larger amounts of source material are available and the criterion becomes the product of efficiency and usable volume of source, the picture changes considerably. Sources of volume as large as 1 cm^3 could probably be used with the diffracting spectrometer in question. With the Pasadena spectrometer operated under the conditions described, the volume of the source is only about 0·03 cm^3. The ritron has been operated with a source volume as large as 12 cm^3 without a serious worsening of the resolution ; the use of this figure assumes that the γ-ray energy is sufficiently large that the absorption in 7 cm of source material can be tolerated. The position of the photoelectric spectrometer can be considerably improved by replacing the lens spectrometer by a large $\pi\sqrt{2}$ spectrometer, though there would be a three-fold diminution in transmission. With the Chalk River $\pi\sqrt{2}$ spectrometer one ought to be able to use a source volume of 10 cm^3 or more, though the fraction of γ-rays entering the converter would be decreased somewhat.

To compare the counting rates of the instruments in figure 6.2 with those of figure 6.1 one must take into account the usable volume of source material. The dimension of the source in the direction of emission in the case of spectrometers using internally produced charged particles is very small, because of the need to avoid too much energy degradation. For the photon-accepting instruments the source thickness is limited by some spectrometer aberration, or by the absorption of the γ-rays in the source. The thickness permitted in a photon-accepting spectrometer may be many orders of magnitude greater than that allowed in an instrument using internally produced charged particles ; the ratio of allowed thickness depends upon the size of the instrument, and varies rapidly with the energy. We have examined the question of relative source thicknesses for spectrometers using internal and external conversion electrons in § 3.1.

The last group of spectrometers which we shall consider consists of those designed to accept photons from a source of large volume. These instruments have been developed to examine the spectra of γ-rays from sources continuously irradiated in the neutron flux of a reactor. The source is necessarily separated from the first stage of the spectrometer by 3 or 4 metres of shielding. Since the diameter of the experimental hole through the shielding is usually only 10 to 20 cm, we are presented with a good approximation to a parallel beam of γ-rays incident on the spectrometer. The spectrometers we are to discuss have been designed, therefore, to accept a broad beam of almost parallel γ-rays ; when necessary, thin absorbers are placed longitudinally in the experimental hole to reduce the divergence

of the beam. The dimension of the source in the direction of the beam can be as large as self-absorption makes practicable, i.e. this dimension is not limited by a spectrometer aberration. We shall assume that the entrance aperture of each of the spectrometers to be compared subtends a solid angle of 3×10^{-5} at the source; this solid angle is consistent with a beam divergence angle of about 1°. In figure 6.3 we give efficiency curves for several spectrometers; the absorption in the source is ignored. Curve *a*, taken from figure 2.11, refers to a two-crystal diffraction spectrometer; at 1 MeV the resolution is 0·5%, at 5 MeV, 0·6%. An experimental curve is also available for a 180° pair spectrometer. In *b* we give the efficiency for a spectrometer with thick slits (see § 5); the resolution is roughly 1% at all energies.

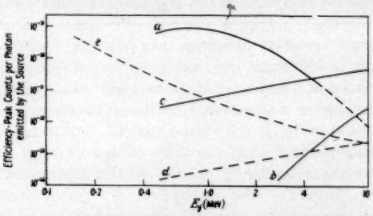


Figure 6.3. Efficiency curves for spectrometers using photons emitted from a large source. Curve *a* refers to a two-crystal diffraction spectrometer, *b* to a flat-field 180° pair spectrometer, and *c* to a Compton spectrometer. A theoretical curve for a Compton spectrometer with better resolution is shown in *d*. The curve *e* shows how a $\pi\sqrt{2}$ spectrometer of the size of the Chalk River instrument would perform in this application.

Curve *c* is based upon the Compton spectrometer of Groshev *et al.* (1956); we have renormalized Lee-Whiting's theoretical curve (see § 4) to the measured efficiency at 1·33 MeV. The resolution for this spectrometer, shown in figure 4.6, is about 2·3% above 2 MeV. To give some idea of the performance of a Compton spectrometer with a resolution comparable with that of the crystal diffraction spectrometer, we have shown in curve *d* the theoretical estimate for Lee-Whiting's design.

Though there is no photoelectric spectrometer designed specially for use with a broad source, the $\pi\sqrt{2}$ type of spectrometer does permit the use of a converter of considerable height. A spectrometer of the size of the Chalk River instrument could tolerate a converter of area 1 cm × 15 cm at a resolution of 1%. If the spectrometer were placed against the shielding of a reactor, the solid angle subtended by the converter at the source would be about $4\pi \times 10^{-5}$. The curve *e* in figure 6.3 shows the efficiency of such a conversion-electron spectrometer, assuming a converter of uranium 8 mg per cm² thick and a corrected transmission equal to $0{\cdot}3 \times 10^{-2}$; these operating conditions should correspond to a resolution of about 1% at 1 MeV.

Although spectrometers using crystal diffraction or electron conversion appear in many cases to be more efficient than those using the Compton process or pair production, much of this advantage disappears when we take the background counting rate into consideration. The fact that the pair and Compton spectrometers count coincidences reduces their backgrounds enormously with respect to spectrometers of the other two types.

Acknowledgments

The authors are much indebted to Drs. R. L. Graham and L. G. Elliott and especially to Dr. H. Carmichael for their careful reading of the manuscript and their many helpful suggestions.

References

Alaga, G., Alder, K., Bohr, A., and Mottelson, B. R., 1955, *K. Danske Vidensk. Selsk., Mat-Fys. Medd.*, **29**, No. 9.

Alburger, D. E., 1952, *Rev. Sci. Instrum.*, **23**, 671 ; 1955, *Beta and Gamma-ray Spectroscopy*, ed. K. Siegbahn (Amsterdam : North-Holland) ; 1956, *Rev. Sci. Instrum.*, **27**, 991.

Alichanov, A. I., and Dzhelepov, V. P., 1938, *C.R. Acad. Sci. U.S.S.R.*, **20**, 113.

Alichanov, A. I., and Latyshev, G. D., 1938, *C.R. Acad. Sci. U.S.S.R.*, **20**, 429.

Balzer, R., Knoepfel, H., Stoll, P., and Wölfli, W., 1958, *Helv. Phys. Acta*, **31**, 328.

Bame, S. J., and Baggett, L. M., 1951, *Phys. Rev.*, **84**, 891.

Bartholomew, G. A., Campion, P. J., and Robinson, K., 1960, *Canad. J. Phys.*, **38**, 194.

Bartholomew, G. A., and Higgs, L. A., 1958, *Atomic Energy of Canada Limited, Report* AECL 669.

Barton, G. W., 1950, *University of California Radiation Laboratory, Report* UCRL-670.

Barton, G. W., Robinson, H. P., and Perlman, I., 1951, *Phys. Rev.*, **81**, 208.

Beckman, O., 1958, *Nuclear Instruments*, **2**, 1.

Beckman, O., and Bergvall, P., 1959, *Phys. Rev.*, **114**, 1280.

Beckman, O., Bergvall, P., and Axelsson, B., 1958, *Ark. Fys.*, **14**, 26.

Beiduk, F. M., and Konopinski, E. J., 1948, *Rev. Sci. Instrum.*, **19**, 594.

Bell, P. R., 1955, *Beta and Gamma-ray Spectroscopy*, ed. K. Siegbahn (Amsterdam : North-Holland).

Bent, R. D., Bonner, T. W., and Sippel, R. F., 1955 a, *Phys. Rev.*, **98**, 1237.

Bent, R. D., Bonner, T. W., McCrary, J. H., Ranken, W. A., and Sippel, R. F., 1955 b, *Phys. Rev.*, **99**, 710.

Bent, R. D., and Kruse, T. H., 1957, *Phys. Rev.*, **108**, 802.

Bergvall, P., 1959, *Ark. Fys.*, **16**, 57.

Bethe, H. A., and Ashkin, J., 1952, *Experimental Nuclear Physics*, Vol. I, ed. E. Segrè (New York : John Wiley).

Bethe, H. A., and Heitler, W., 1934, *Proc. Roy. Soc.* A, **146**, 83.

Bishop, G., and Wilson, R., 1955, *Beta and Gamma-ray Spectroscopy*, ed. K. Siegbahn (Amsterdam : North-Holland).

Blewett, J. P., 1947, *J. Appl. Phys.*, **18**, 968.

Bohr, A., and Mottelson, B. R., 1955, *Beta and Gamma-ray Spectroscopy*, ed. K. Siegbahn (Amsterdam : North-Holland).

Browne, C. I., 1952, *University of California Radiation Laboratory, Report* UCRL-1760.

Cauchois, Y., 1932, *C.R. Acad. Sci., Paris*, **195**, 1479 ; 1952, *J. Phys. Radium*, **13**, 113 ; 1955, *Ibid.*, **16**, 253.

Chubinskii, O. V., 1957, *Izvest. Akad. Nauk S.S.S.R., Ser. Fiz.*, **21**, 1583.

Chupp, E. L., Clark, A. F., DuMond, J. W. M., Gordon, F. J., and Mark, H., 1957 a, *Phys. Rev.*, **107**, 745 ; 1957 b, *Bull. Amer. Phys. Soc.*, [II] **2**, 396 ; 1958, *Ibid.*, **3**, 55.

Chupp, E. L., DuMond, J. W. M., and Mark, H., 1958, *Rev. Sci. Instrum.*, **29**, 1153.

Chupp, E. L., DuMond, J. W. M., Sewell, R., and Mark, H., 1959, *Bull. Amer. Phys. Soc.*, [II], **4**, 141.

Cohen, L. R., DuMond, J. W. M., Layton, T., and Rollett, R. S., 1955, *Rev. Mod. Phys.*, **27**, 363.

Crowe, K. M., and Phillips, R. H., 1954, *Phys. Rev.*, **96**, 470.

Daniel, H., and Bothe, W., 1954, *Z. Naturforsch.*, **9a**, 402.

Danysz, J., 1912, *Le Radium*, **9**, 1.

Davis, R. C., Bell, P. R., Kelley, G. G., and Lazar, N. H., 1956, *Inst. Radio Engrs Trans. Nucl. Sci.*, **3**, No. 4, 82.

Davisson, C. M., and Evans, R. D., 1952, *Rev. Mod. Phys.*, **24**, 79.

DAY, P. P., 1955, *Phys. Rev.*, **97**, 689.
DUMOND, J. W. M., 1933, *Rev. Mod. Phys.*, **5**, 1 ; 1947, *Rev. Sci. Instrum.*, **18**, 626 ; 1949, *Ibid.*, **20**, 160 ; 1955 a, *Ergebn. der exakt. Naturwiss.*, **28**, 232 ; 1955 b, *Beta and Gamma-ray Spectroscopy*, ed. K. Siegbahn (Amsterdam : North-Holland), Chap. IV ; 1957, *Annals of Physics*, **2**, 283 ; 1958 a, *Proc. Rehovoth Conf. Nucl. Structure*, ed. H. A. Lipkin (Amsterdam : North-Holland), p. 536 ; 1958 b, *Ann. Rev. Nucl. Sci.*, **8**, 163 ; 1959, *Proc. U.S. Nat. Acad. Sci.*, **45**, 1052.
DUMOND, J. W. M., and KIRKPATRICK, H. A., 1930, *Rev. Sci. Instrum.*, **1**, 90.
DZHELEPOV, B. S., 1951, *Dokl. Akad. Nauk S.S.S.R.*, **77**, 389 ; 1955, *Ibid.*, **101**, 825 ; 1957, *Izvest. Akad. Nauk S.S.S.R., Ser. Fiz.*, **21**, 1580.
DZHELEPOV, B. S., and BASHILOV, A. A., 1950, *Izvest. Akad. Nauk S.S.S.R., Ser. Fiz.*, **14**, 263.
DZHELEPOV, B. S., and ORBELI, M. L., 1948, *Doklady Akad. Nauk S.S.S.R.*, **62**, 615. English translation NRC-TT-630.
DZHELEPOV, B. S., PRIKHODTSEVA, V. P., FEOKTISTOV, A. I., and KHOL'NOV, IU. V., 1956, *Izvest. Akad. Nauk S.S.S.R., Ser. Fiz.*, **20**, 1361 (Translation, *Bull. Acad. Sci. U.S.S.R., Phys. Ser.*, **20**, 1245).
DZHELEPOV, B. S., and SHESTOPALOVA, S. Z., 1956, *Izvest. Akad. Nauk S.S.S.R., Ser. Fiz.*, **20**, 328 (Translation, *Bull. Acad. Sci. U.S.S.R., Phys. Ser.*, **20**, 302).
DZHELEPOV, B. S., ZHUKOVSKI, N. N., and KHOL'NOV, IU. V., 1954, *Izvest. Akad. Nauk S.S.S.R., Ser. Fiz.*, **18**, 599 (Translation, *Bull. Acad. Sci. U.S.S.R., Phys. Ser.*, **18**, 285).
FRILLEY, M., GOKHALE, B. G., and VALADARES, M., 1951, *C.R. Acad. Sci., Paris*, **232**, 50.
GAERTTNER, E. R., and PARDUE, L. A., 1940, *Phys. Rev.*, **57**, 386.
GIBSON, W. M., GROTDAL, T., ORLIN, J. J., and TRUMPY, B., 1951, *Phil. Mag.*, **42**, 555.
GLAZUNOV, M. P., and SPITSYN, V. I., 1956, *Izvest. Akad. Nauk S.S.S.R., Ser. Fiz.*, **20**, 1423 (Translation, *Bull. Acad. Sci. U.S.S.R., Phys. Ser.*, **20**, 1302).
GOLDSTEIN, S., and TALMI, I., 1957, *Phys. Rev.*, **105**, 995.
GROSHEV, L. V., ADYASEVICH, B. P., and DEMIDOV, A. M., 1956, *Proc. Intern. Conf. on Peaceful Uses of Atomic Energy, Geneva*, **2**, 39.
HAMERMESH, B., ROSE, D., SMITHER, R. K., 1956, *Bull. Amer. Phys. Soc.*, [11] **1**, 346.
HEDGRAN, A., 1951, *Manne Siegbahn*, ed. Melander, L., Sievert, R., Slätis, H., and Westgren, A. (Uppsala : Almqvist and Wiksells).
HOWLAND, P. R., SCOFIELD, N. E., and TAYLOR, R. A., 1956, *Nucleonics*, **14**, No. 6, 50.
KERMAN, A. K., 1956, *K. Danske Vidensk. Selsk., Mat.-Fys. Medd.*, **30**, No. 15.
KINSEY, B. B., and BARTHOLOMEW, G. A., 1953, *Canad. J. Phys.*, **31**, 537.
KINSEY, B. B., BARTHOLOMEW, G. A., and WALKER, W. H., 1951, *Phys. Rev.*, **83**, 519.
KNOWLES, J. W., 1957, *Atomic Energy of Canada Limited, Chalk River Project Report*, GPI-42 ; 1959, *Canad. J. Phys.*, **37**, 203.
KORSUNSKY, M., KELMAN, V., and PETROV, B., 1945, *J. Phys. U.S.S.R.*, **9**, 7.
LATYSHEV, G. D., 1947, *Rev. Mod. Phys.*, **19**, 132.
LEE-WHITING, G. E., 1957, *Atomic Energy of Canada Limited Report*, CRT-673 ; 1960, *Canad. J. Phys.*, in the press.
LEE-WHITING, G. E., and TAYLOR, E. A., 1957, *Canad. J. Phys.*, **35**, 1, and *Atomic Energy of Canada Limited Report*, CRT-668.
MCDANIEL, B. D., VON DARDEL, G., and WALKER, R. L., 1947, *Phys. Rev.*, **72**, 985.
MALMFORS, K. G., 1958, *Ark. Fys.*, **13**, 237.
MERRILL, J. J., and DUMOND, J. W. M., 1958, *Phys. Rev.*, **110**, 79.
MLADJENOVIC, M., and HEDGRAN, A., 1952, *Physica*, **18**, 1242.
MOAK, C. D., and DABBS, J. W. T., 1949, *Phys. Rev.*, **75**, 1770.
MOTZ, H. T., 1956, *Phys. Rev.*, **104**, 1353.
MOTZ, H. T., CARTER, R. E., and FISHER, P. C., 1959, *Bull. Amer. Phys. Soc.*, II, **4**, 246.
MOTZ, J. W., MILLER, W., WYCKOFF, H. O., GIBSON, H. F., and KIRN, F. S., 1953, *Rev. Sci. Instrum.*, **24**, 929.
MULLER, D. E., HOYT, H. C., KLEIN, D. J., and DUMOND, J. W. M., 1952, *Phys. Rev.*, **88**, 775.
NELMS, A. T., 1953, U.S. Dept. of Commerce, N.B.S. Circular 542.
PAVINSKI, P. P., 1954, *Izvest. Akad. Nauk S.S.S.R., Ser. Fiz.*, **18**, 175.

PERSICO, E., 1949, *Rev. Sci. Instrum.*, **20**, 191.
RAE, E. R., RUTHERGLEN, J. G., and SMITH, R. D., 1950, *Proc. Phys. Soc.* A, **63**, 775; 1951, *Ibid.*, **64**, 906.
RANKEN, W. A., BONNER, T. W., MCCRARY, J. M., and RABSON, T. A., 1958, *Phys. Rev.*, **109**, 917.
RICHARDSON, J. R., and EMO, L., 1938, *Phys. Rev.*, **53**, 234.
ROSE, D., OSTRANDER, H., and HAMERMESH, B., 1957, *Rev. Sci. Instrum.*, **28**, 233.
ROSE, M. E., 1949, *Phys. Rev.*, **76**, 678; 1958, *Internal Conversion Coefficients* (Amsterdam: North-Holland).
ROSE, M. E., GOERTZEL, G. H., SPINRAD, B. I., HARR, J., and STRONG, P., 1951, *Phys. Rev.*, **83**, 79.
RYDE, N., and ANDERSSON, B., 1955, *Proc. Phys. Soc.* B, **68**, 12, 1117.
SERSON, P. H., 1957, *Canad. J. Phys.*, **35**, 1387.
SIEGBAHN, K. (ed.), 1955, *Beta and Gamma-ray Spectroscopy* (Amsterdam: North-Holland).
SIEGBAHN, K., and JOHANSSON, S., 1950, *Rev. Sci. Instrum.*, **21**, 442.
SLÄTIS, H., and SIEGBAHN, K., 1949, *Arkiv Fysik.*, **1**, 339.
SUMBAEV, O. I., 1957 a, *Zhur. Eksp. i Teoret. Fiz.*, **32**, 247 (Translation, *Soviet Physica J.E.T.P.*, **5**, 170); 1957 b, *Ibid.*, **32**, 1276 (Translation, *Ibid.*, **5**, 1042).
SVARTHOLM, N., 1946, *Arkiv Mat. Astron. Fysik*, **33**A, No. 24.
SVARTHOLM, N., and SIEGBAHN, K., 1946, *Arkiv Mat. Astron. Fysik*, **33**A, No. 21.
TALMI, I., and THIEBERGER, R., 1956, *Phys. Rev.*, **103**, 718.
TERRELL, J., 1950, *Phys. Rev.*, **80**, 1076.
TERRELL, J., and PHILLIPS, G. C., 1951, *Phys. Rev.*, **83**, 703.
THOMAS, R. G., and LAURITSEN, T., 1952, *Phys. Rev.*, **88**, 969.
ULMER, K., and ZIEGLER, B., 1956, *Z. angew. Phys.*, **8**, 49.
VERSTER, N. F., 1950, *Physica*, **16**, 815.
WALKER, R. L., and MCDANIEL, B. D., 1948, *Phys. Rev.*, **74**, 315.
WAPSTRA, A. H., NIJGH, G. J., and VAN LIESHOUT, R., 1959, *Nuclear Spectroscopy Tables* (Amsterdam: North-Holland).
ZACHARIASEN, W. H., 1946, *Theory of X-ray Diffraction in Crystals* (New York: John Wiley).
ZOLOTAVIN, A. V., 1954, *Izvest. Akad. Nauk S.S.S.R. Ser. Fiz.*, **18**, 127.
ZOLOTAVIN, A. V., and SADOVSKI, V. S., 1954, *Izvest. Akad. Nauk., S.S.S.R., Ser. Fiz.*, **18**, 215 (English translation, *Atomic Energy of Canada Limited Report*, No. 181).
ZWORYKIN, V. K., MORTON, G. A., RAMBERG, E. G., HILLIER, J., and VANCE, A. W., 1945, *Electron Optics and the Electron Microscope* (New York: John Wiley).

^{3}He INDUCED REACTIONS

By D. A. BROMLEY and E. ALMQVIST
Atomic Energy of Canada, Ltd., Chalk River, Ontario

CONTENTS

PAGE

§ 1. Introduction 545
§ 2. Reaction mechanisms 552
2.1. Introduction 552
2.2. Compound system mechanisms 553
2.2.1. Resonance region 553
2.2.2. Intermediate region 557
2.2.3. Non-resonant region 559
2.3. Direct interaction mechanisms 561
2.3.1. Single nucleon transfer 561
2.3.1.1. (^{3}He, α) reactions 563
2.3.1.2. (^{3}He, d) reactions 567
2.3.1.3. (^{3}He, 2p) and (^{3}He, np) reactions 568
2.3.2. Two-nucleon transfer 569
2.3.2.1. (^{3}He, p), (^{3}He, n) and (^{3}He, t) reactions —survey 569
2.3.2.2. Final state statistical factors 570
2.3.2.3. Angular distribution calculations 571
2.3.3. Direct surface excitation 575
2.3.4. Knock-on processes 576
2.3.5. Summary—direct interaction 577
2.4. Summary—reaction mechanisms 577
§ 3. Nuclear spectroscopy 578
3.1. Introduction 578
3.2. Particle spectroscopy 578
3.3. Determination of level characteristics 580
3.4. Coulomb excitation measurements 585
§ 4. Survey of experimental results 587
4.1. Average reaction cross sections 587
4.2. ^{2}H target 588
4.3. ^{3}H target 588
4.4. ^{3}He target 589
4.5. ^{4}He target 589
4.6. ^{6}Li target 592
4.7. ^{7}Li target 595
4.8. ^{9}Be target 598
4.9. ^{10}B target 600
4.10. ^{11}B target 604
4.11. ^{12}C target 605
4.12. ^{13}C target 609
4.13. ^{14}N target 610
4.14. ^{16}O target 612
4.15. ^{19}F target 614
4.16. ^{20}Ne target 614
4.17. ^{24}Mg target 615
4.18. ^{28}Si target 615
4.19. ^{40}Ca target 615
4.20. Higher A targets 615

PAGE

§ 5. Suggested experiments 616
5.1. General comments 616
5.2. Nuclear spectroscopic investigations 616
5.3. Angular distribution and correlation measurements 617
5.4. Isobaric multiplet studies 617
5.5. Neutron measurements 618
5.6. Lifetime measurements 618
5.7. Shell effects 618
5.8. Coulomb excitation measurements 618
5.9. Nucleon cluster phenomena 619
§ 6. Conclusions 619
Acknowledgments 620
References 620
Appendix : Index of additional topics reviewed in Chalk River report CRP-881 628

Abstract. This report presents a review of the results obtained from study of ^{3}He induced nuclear reactions prior to June 1959. Following a brief historical survey including range–energy, energy loss, and stopping power data as well as a reaction Q value tabulation, the experimental results are first presented in terms of their relevance to determination of the reaction mechanisms involved ; this is followed by a discussion of the use of ^{3}He in nuclear spectroscopy. The next section is devoted to summarizing available ^{3}He reaction data according to target nuclide involved. Numerous figures illustrating excitation curves, spectra and angular distributions are included. The report concludes with a short section devoted to suggested experiments which emphasizes the very great scope remaining for experimental work with low-energy ^{3}He accelerators. A comprehensive bibliography covering all publications on ^{3}He reactions prior to June 1959, and including selected papers on experimental techniques as well as on the production and handling of ^{3}He in accelerators, is appended.

§ 1. Introduction

Of the light nuclides currently in use as nuclear projectiles, ^{3}He has the unique distinction of having been first conclusively identified through its induction of a nuclear reaction. Alvarez and Cornog in 1939, after adjusting the Crocker cyclotron at the University of California for acceleration of a postulated helium isotope of mass 3, identified the characteristic β^+ decay of ^{30}P produced in the ^{28}Si(^{3}He, p)^{30}P reaction on a quartz target.

It was immediately obvious that the ^{3}He ion had marked advantages for use in nuclear studies ; these advantages follow directly from the first two ^{3}He characteristics listed in table 1.

Table 1. ^{3}He Characteristics

Mass excess	15·814 MeV
Charge to mass ratio	2/3
Natural isotopic abundance	10^{-6} atmospheric helium
	10^{-7} well helium

As a result of the high mass excess, hence high ^{3}He binding energies in nuclear reaction compound systems, ^{3}He induced reactions are characterized by high Q values and by highly excited compound systems. Figure 1 is a graphical presentation of the compound system binding energies for ^{3}He reactions on available targets for $A \leqslant 40$. This figure and table 2, listing the Q values for these reactions,

are taken from Bromley and Rutledge (1958) and are based on the atomic masses tabulated by Mattauch *et al.* (1956 a), Wapstra (1955 a, b), and where otherwise unavailable, on masses calculated from mirror pair extrapolations (Cameron 1958, private communication). In figure 1 the median binding energy is some 17 MeV; the various symbols identify the (N, Z) families in the compound systems and demonstrate the expected pairing energy shifts. In effect, when used in conjunction with a low energy ($\lesssim 5$ MeV) accelerator, these ions lead to nuclear excitations which would otherwise be inaccessible except with much higher energy accelerators. This advantage of using low energy accelerators of course can be exploited only for elements that have a sufficiently low Coulomb barrier to give an observable reaction

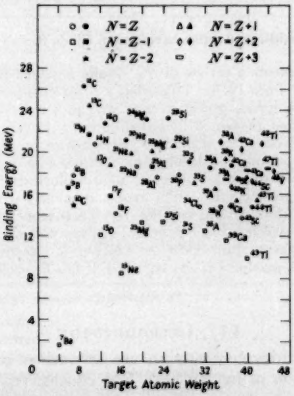


Figure 1. Compound system binding energies in ^{3}He induced reactions. Solid symbols refer to compound systems with even N, open symbols with odd N. From Bromley and Rutledge (1958).

yield. It should be noted that because ^{3}He ions have both larger radius and mass than nucleons they can impart more angular momentum to the compound system than a nucleon of the same energy. Consequently the probability of p-wave or higher angular momentum interactions relative to s-wave is greater for ^{3}He than for nucleons. Figure 2 presents the s-wave ^{3}He Coulomb barrier penetrabilities for representative energies as functions of target atomic number. For convenience, stopping-power and range–energy data for ^{3}He ions in representative materials throughout the atomic table are shown in figures 3, 4, 5, 6 and 7. These data have been prepared from those tabulated by Whaling (1958).

The high charge to mass ratio permits the study of the proton-rich members of isobaric multiplets which have until recently received very little study because of their relative inaccessibility to precision techniques. Of particular interest is the fact that (^{3}He, p) and (^{3}He, n) reactions lead simultaneously to members of isobaric multiplets and thus give direct information regarding the charge independence of nuclear forces without obvious complications from differing reaction mechanisms.

In elucidating the mechanisms for nuclear reactions, (^{3}He, α) and (^{3}He, d) reactions are of particular interest in that they involve transfer of a single nucleon and are thus analogous to the (d, t) and (d, n) reactions; these latter have received extensive study following the publication of Butler's (1951) pioneering study on the interpretation of such reactions in terms of a direct stripping or pick-up interaction. In recent years much theoretical effort has been expended in evolving a more detailed and precise understanding of these reactions involving nuclear and Coulomb corrections. Because of technical difficulties, the (^{3}He, α) and (^{3}He, d) reactions have received relatively little study as yet; however, when data on these

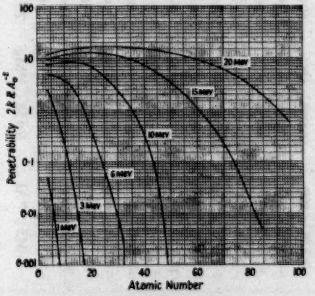


Figure 2. S wave ^{3}He penetrabilities. In preparing these curves typical nuclei, e.g. ^{9}Be, ^{16}O, ^{24}Mg, ^{31}P, ^{40}Ca, ^{56}Fe, ^{90}Zr, ^{148}Nd, ^{204}Pb and ^{238}U, have been used. Coulomb functions have been obtained from Tubis (1958) and from Sharp, Gove and Paul (1953).

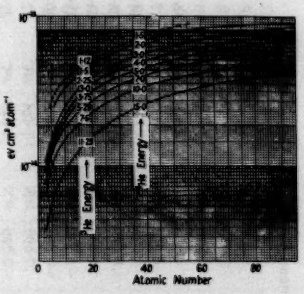


Figure 3. ^{3}He energy loss in ev per atom/cm^2 as a function of atomic number and ^{3}He incident energy. Prepared from the compilation of Whaling (1958).

reactions become available in quantity, comparison of these with the corresponding data from the deuteron induced reactions will provide direct information on these nuclear and Coulomb corrections.

^{3}He can be used for Coulomb excitation studies but in general is not as suitable for this as is ^{4}He because the high charge to mass ratio results in a relatively high bremsstrahlung yield and reduced Coulomb excitation yield. However, it is possible to determine a gamma-ray transition multipolarity by measuring the ratio of the Coulomb excitation cross section for ^{3}He to that for ^{4}He incident ions (Bjerregaard and Huus 1954) and for this reason ^{3}He Coulomb excitation has been studied (Bromley, Kuehner and Almqvist 1959 a). In evaluating the relative cross sections with sufficient accuracy for application of this technique it is advantageous to use particles such as ^{3}He and ^{4}He for which the range–energy relations in the target material can be accurately related.

Barkas (1939) first drew attention to some features which make ^{3}He ions useful in reaction studies. However, the low natural isotopic abundance of ^{3}He precluded the possibility of detailed reaction measurements using naturally occurring helium and systematic ^{3}He studies awaited the availability of isotopically enriched ^{3}He.

Table 2. ^{3}He Induced Reaction Q Values

Target \ Final	n	p	d	t	α	γ	nn
^{4}He		−4·023 ; ^{6}Li	−7·256 ; ^{5}Li		0 ; ^{5}He	1·584 ; ^{7}Be	
^{6}Li	−1·616 ; ^{8}B	16·779 ; ^{8}Be	0·113 ; ^{7}Be		15·104 ; ^{5}Li	16·594 ; ^{9}B	
^{7}Li	9·342 ; ^{9}B	11·194 ; ^{9}Be	11·753 ; ^{8}Be	−0·882 ; ^{7}Be	13·320 ; ^{6}Li	17·780 ; ^{10}B	−8·868 ; ^{8}B
^{7}Be		10·987 ; ^{9}B	−4·997 ; ^{8}B			14·892 ; ^{10}C	
^{9}Be	7·566 ; ^{11}C	10·330 ; ^{11}B	1·092 ; ^{10}B	−1·089 ; ^{9}B	18·905 ; ^{6}Be	26·287 ; ^{15}C	−5·667 ; ^{10}C
^{10}Be	19·474 ; ^{12}C	6·880 ; ^{12}B	5·743 ; ^{11}B	0·536 ; ^{10}B	13·759 ; ^{9}Be	24·421 ; ^{13}C	0·753 ; ^{11}C
^{10}B	1·248 ; ^{13}N	19·701 ; ^{12}C	3·206 ; ^{11}C	−3·770 ; ^{10}C	12·134 ; ^{9}B	21·645 ; ^{13}N	
^{11}B	10·181 ; ^{13}N	13·184 ; ^{13}C	10·463 ; ^{12}C	−2·001 ; ^{11}C	9·108 ; ^{10}B	20·724 ; ^{14}N	−10·216 ; ^{12}N
^{12}C	−1·167 ; ^{14}O	4·767 ; ^{14}N	−3·550 ; ^{13}N	−17·690 ; ^{13}N	1·401 ; ^{11}C	12·122 ; ^{15}O	
^{13}C	7·175 ; ^{15}O	10·663 ; ^{15}N	2·046 ; ^{14}N	−2·240 ; ^{13}N	15·625 ; ^{12}C	22·771 ; ^{16}O	−6·114 ; ^{14}O
^{14}C	14·604 ; ^{16}O	5·146 ; ^{16}N	4·722 ; ^{15}N	0·136 ; ^{14}N	12·405 ; ^{13}C	18·750 ; ^{17}O	−0·992 ; ^{15}O
^{14}N	−1·129 ; ^{16}F	15·231 ; ^{16}O	1·861 ; ^{15}O	−5·171 ; ^{14}O	10·029 ; ^{12}N	15·826 ; ^{17}F	−14·580 ; ^{15}F
^{15}N	4·982 ; ^{17}F	8·534 ; ^{17}O	6·614 ; ^{16}O	−2·725 ; ^{15}O	9·729 ; ^{14}N	14·161 ; ^{18}F	−11·973 ; ^{16}F
^{16}O	−2·967 ; ^{18}Ne	2·053 ; ^{18}F	−4·899 ; ^{17}F	−15·597 ; ^{16}F	4·976 ; ^{15}O	8·416 ; ^{19}Ne	−22·294 ; ^{17}Ne
^{17}O	4·270 ; ^{19}Ne	8·308 ; ^{19}F	0·133 ; ^{18}F	−2·788 ; ^{17}F	16·426 ; ^{16}O	21·186 ; ^{20}Ne	−7·113 ; ^{18}Ne
^{18}O	13·105 ; ^{20}Ne	6·830 ; ^{20}F	2·453 ; ^{19}F	−1·691 ; ^{18}F	12·491 ; ^{17}O	19·859 ; ^{21}Ne	−3·811 ; ^{19}Ne
^{19}F	7·603 ; ^{21}Na	11·912 ; ^{21}Ne	7·383 ; ^{20}Ne	−3·275 ; ^{19}Ne	10·171 ; ^{18}F	18·655 ; ^{22}Na	−10·963 ; ^{20}Na
^{20}Ne	0·273 ; ^{22}Mg	5·777 ; ^{22}Na	−3·049 ; ^{21}Na	−15·358 ; ^{20}Na	3·656 ; ^{19}Ne	13·319 ; ^{23}Mg	−18·959 ; ^{21}Mg
^{21}Ne	6·565 ; ^{23}Mg	11·440 ; ^{23}Na	1·249 ; ^{22}Na	−3·546 ; ^{21}Na	13·818 ; ^{20}Ne	23·129 ; ^{24}Mg	−6·481 ; ^{22}Mg
^{22}Ne	12·764 ; ^{24}Mg	8·032 ; ^{24}Na	3·301 ; ^{23}Na	−2·859 ; ^{22}Na	10·207 ; ^{21}Ne	20·097 ; ^{25}Mg	−3·800 ; ^{23}Mg
^{23}Na	1·588 ; ^{24}Al	16·386 ; ^{24}Mg	2·048 ; ^{23}Mg	−4·741 ; ^{23}Mg	9·520 ; ^{21}Na	18·645 ; ^{25}Al	−13·224 ; ^{23}Al
^{23}Na	6·228 ; ^{25}Al	11·302 ; ^{25}Mg	6·195 ; ^{24}Mg	−4·112 ; ^{23}Mg	8·155 ; ^{22}Na	17·608 ; ^{26}Al	−10·829 ; ^{24}Al
^{24}Mg	−0·064 ; ^{26}Si	5·919 ; ^{26}Al	−3·235 ; ^{25}Al	−14·035 ; ^{24}Al	4·008 ; ^{23}Mg	13·370 ; ^{27}Si	−18·879 ; ^{25}Si
^{25}Mg	6·037 ; ^{27}Si	11·644 ; ^{27}Al	0·812 ; ^{26}Al	−4·311 ; ^{25}Al	13·239 ; ^{24}Mg	23·235 ; ^{28}Si	−7·397 ; ^{26}Si
^{26}Mg	12·121 ; ^{28}Si	8·254 ; ^{28}Al	2·756 ; ^{27}Al	−4·045 ; ^{26}Al	9·458 ; ^{25}Mg	20·593 ; ^{29}Si	−5·077 ; ^{27}Si

^{27}Al	6·598 ;	^{29}P	12·343 ;	^{28}Si	6·097 ;	^{28}Si	−4·844 ;	^{27}Si	7·514 ;	^{26}Al	17·856 ;	^{30}P	−10·674 ;	^{28}P
^{28}Si	−0·446 ;	^{30}S	6·265 ;	^{30}P	−2·767 ;	^{29}P	−13·782 ;	^{28}P	3·374 ;	^{27}Si	12·437 ;	^{31}S	−19·453 ;	^{29}S
^{28}Si	3·965 ;	^{31}S	10·184 ;	^{31}P	0·019 ;	^{30}P	−4·982 ;	^{29}P	12·100 ;	^{28}Si	19·046 ;	^{32}S	−8·918 ;	^{30}S
^{30}Si	8·437 ;	^{32}S	7·510 ;	^{32}P	1·801 ;	^{31}P	−4·333 ;	^{30}P	9·963 ;	^{29}Si	17·083 ;	^{33}Si	−6·644 ;	^{31}S
^{31}P	3·557 ;	^{33}Cl	9·788 ;	^{33}S	3·368 ;	^{32}S	−5·456 ;	^{31}S	8·181 ;	^{30}P	14·879 ;	^{34}Cl	−12·670 ;	^{32}Cl
^{32}S	−0·074 ;	^{34}A	6·017 ;	^{34}Cl	−3·079 ;	^{33}Cl	−13·049 ;	^{32}Cl	5·491 ;	^{31}S	12·535 ;	^{35}A	−16·538 ;	^{33}A
^{34}S	7·716 ;	^{36}A	7·627 ;	^{36}Cl	1·289 ;	^{35}Cl	−5·011 ;	^{34}Cl	9·707 ;	^{33}S	16·552 ;	^{37}A	−6·976 ;	^{35}A
^{35}S	9·154 ;	^{37}A	10·752 ;	^{37}Cl	2·455 ;	^{36}Cl	0·148 ;	^{35}Cl	13·174 ;	^{34}S	20·897 ;	^{38}A	0·318 ;	^{36}A
^{36}S	10·847 ;	^{38}A	6·810 ;	^{38}Cl	2·928 ;	^{37}Cl	−1·338 ;	^{36}Cl	10·522 ;	^{35}S	17·486 ;	^{39}A	−0·896 ;	^{37}A
^{35}Cl	2·856 ;	^{37}K	9·769 ;	^{37}A	3·159 ;	^{36}A	−5·276 ;	^{35}A	8·015 ;	^{34}Cl	14·854 ;	^{38}K	−12·204 ;	^{36}K
^{36}Cl	6·290 ;	^{38}K	12·948 ;	^{38}A	3·431 ;	^{37}A	0·852 ;	^{36}A	12·008 ;	^{35}Cl	19·370 ;	^{39}K	−5·708 ;	^{37}K
^{37}Cl	8·847 ;	^{39}K	9·064 ;	^{39}A	4·651 ;	^{38}A	−0·835 ;	^{37}A	10·049 ;	^{36}Cl	16·640 ;	^{40}K	−4·233 ;	^{38}K
^{36}A	1·849 ;	^{38}Ca	6·201 ;	^{38}K	−3·571 ;	^{37}K	−12·374 ;	^{36}K	5·880 ;	^{35}A	11·643 ;	^{39}Ca	−13·496 ;	^{37}Ca
^{37}A	2·807 ;	^{38}Ca	10·445 ;	^{39}K	−0·409 ;	^{38}K	−6·150 ;	^{37}K	11·736 ;	^{36}A	18·781 ;	^{40}Ca	−6·987 ;	^{38}Ca
^{38}A	7·038 ;	^{40}Ca	6·495 ;	^{40}K	0·928 ;	^{39}K	−5·895 ;	^{38}K	8·829 ;	^{37}A	15·404 ;	^{41}Ca	−8·936 ;	^{39}Ca
^{38}A	8·765 ;	^{41}Ca	9·985 ;	^{41}K	2·082 ;	^{40}K	0·546 ;	^{39}K	13·933 ;	^{38}A	20·241 ;	^{42}Ca	0·399 ;	^{40}Ca
^{39}K	2·240 ;	^{42}Sc	8·982 ;	^{41}Ca	2·842 ;	^{40}Ca	−6·875 ;	^{39}Ca	7·492 ;	^{38}K	13·798 ;	^{43}Sc	−14·152 ;	^{40}Sc
^{40}K	6·005 ;	^{42}Sc	12·665 ;	^{42}Ca	3·415 ;	^{41}Ca	1·306 ;	^{40}Ca	12·779 ;	^{39}K	17·604 ;	^{43}Sc	−5·553 ;	^{41}Sc
^{41}K	7·475 ;	^{43}Sc	10·474 ;	^{43}Ca	4·762 ;	^{42}Ca	−0·457 ;	^{41}Ca	10·443 ;	^{40}K	17·211 ;	^{44}Sc	−4·124 ;	^{42}Sc
^{40}A	10·391 ;	^{42}Ca	7·504 ;	^{42}K	2·361 ;	^{41}K	−1·511 ;	^{40}K	10·722 ;	^{39}A	18·329 ;	^{43}Ca	−1·085 ;	^{41}Ca
^{40}Ca	−1·768 ;	^{42}Ti	5·462 ;	^{43}Sc	−3·870 ;	^{41}Sc	−14·005 ;	^{40}Sc	4·598 ;	^{39}Ca	9·941 ;	^{43}Ti	−15·544 ;	^{41}Ti
^{41}Ca	1·575 ;	^{43}Ti	8·695 ;	^{43}Sc	−0·678 ;	^{42}Sc	−5·979 ;	^{41}Sc	12·206 ;	^{40}Ca	14·833 ;	^{44}Ti	−10·134 ;	^{42}Ti
^{42}Ca	3·357 ;	^{44}Ti	6·955 ;	^{44}Sc	−0·555 ;	^{43}Sc	−5·897 ;	^{42}Sc	9·096 ;	^{41}Ca	15·459 ;	^{45}Ti	−9·901 ;	^{43}Ti
^{42}Ca	7·521 ;	^{45}Ti	10·348 ;	^{45}Sc	1·243 ;	^{44}Sc	−2·236 ;	^{43}Sc	12·634 ;	^{42}Ca	20·838 ;	^{46}Ti	−4·581 ;	^{44}Ti
^{44}Ca	9·669 ;	^{46}Ti	8·090 ;	^{46}Sc	1·405 ;	^{45}Sc	−3·669 ;	^{44}Sc	9·403 ;	^{43}Ca	18·458 ;	^{47}Ti	−3·648 ;	^{45}Ti
^{45}Sc	7·865 ;	^{47}V	11·559 ;	^{47}Ti	4·996 ;	^{46}Ti	−2·064 ;	^{45}Ti	9·241 ;	^{44}Sc	18·256 ;	^{48}V	−5·411 ;	^{46}V

Various techniques (Daunt 1952, Bowring and Davies 1958, Furry and Jones 1946, and Clusius and Dickel 1938, 1939) have been developed to obtain from natural gas ^{3}He concentrations up to 10%. More recently the advent of high power nuclear reactors makes possible the production, in quantity, of 100% ^{3}He from the decay of tritium ; the price of ^{3}He has dropped from about \$250 000 per litre in

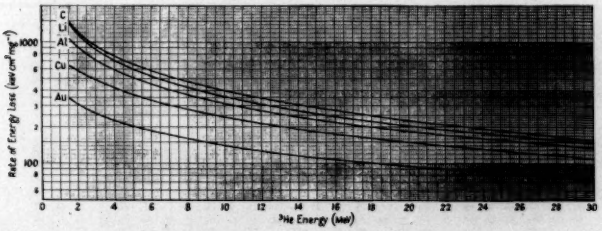


Figure 4. Rate of energy loss for ^{3}He ions in solids. These curves have been prepared from the compilation of stopping cross section data of Whaling (1958).

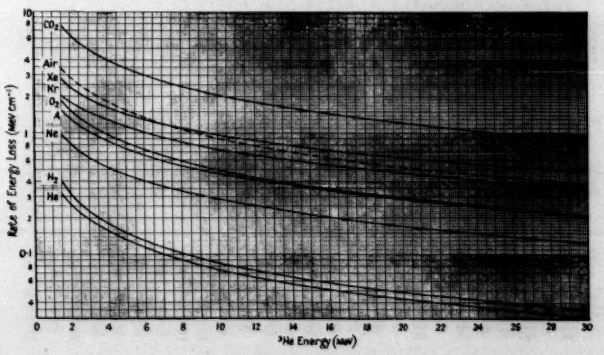


Figure 5. Rate of energy loss for ^{3}He ions in gases. These curves have been prepared from the compilation of stopping cross section data of Whaling (1958).

1950 to \$1500 per litre in 1959. This price is sufficiently low that gas recovery is not essential ; however, a number of laboratories have equipped accelerators with recovery systems to permit extended and economic operation with limited quantities of ^{3}He gas.

It is not possible, within the scope of this article, to consider, in detail, the operation of ^{3}He handling systems (Allen and Almqvist 1953, Kundu *et al.* 1953, Almqvist 1954, Ise *et al.* 1954, Effat and Fremlin 1955, Wegner and Hall 1958) nor to discuss the question of their necessity or desirability. Similarly, it has

not been possible to include a discussion of targetry problems peculiar to the use of ^3He ions or of reaction product detectors, detector systems and measurement

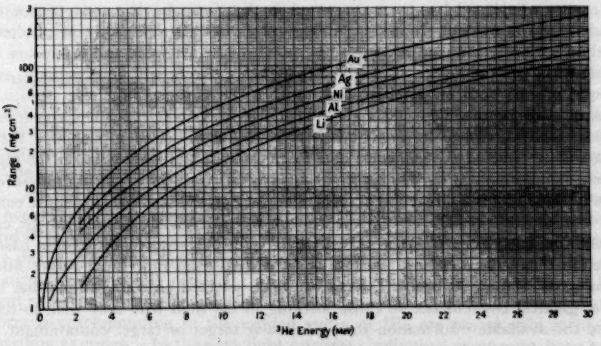


Figure 6. Range–energy curves for ^3He ions in solids. These curves have been prepared from the stopping cross section data presented in figure 4.

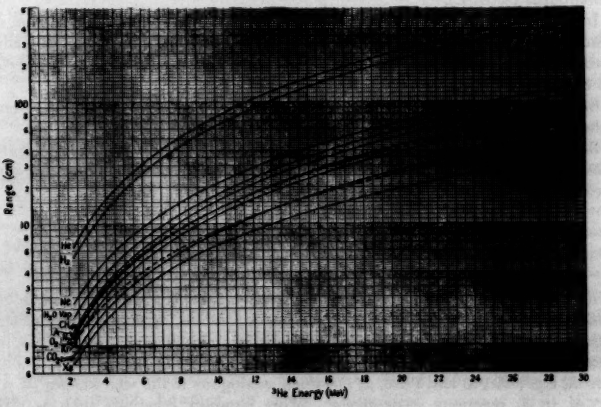


Figure 7. Range–energy curves for ^3He ions in gases. These curves have been prepared from the stopping cross section data of figure 5.

techniques. The reader is referred to a more extensive review article (Bromley and Almqvist 1959), available as Chalk River Report CRP–881, wherein a relatively complete treatment of these topics is presented. Appendix A presents the index

of the sections which have been deleted from the condensed review presented herein.

The bibliography following this review is that corresponding to the more extensive article and has been presented in that form for the convenience of the reader. As is evident from this bibliography, an extensive body of publications dealing with ^{3}He induced reactions has now become available and it appears worth while to survey the field. It is hoped that such a survey will be of value to experimental groups both engaged in, and about to be engaged in, these studies. An attempt will be made to emphasize particular areas where further research may appear fruitful, and to summarize briefly what has been accomplished. We shall first survey selected experimental data in an attempt to throw light on the reaction mechanisms involved in ^{3}He reactions as distinct from the spectroscopic information provided by these reaction studies. Having completed this survey we examine the experimental results first in the light of a broad division based on the energy of the incident ^{3}He ion, since this division is in part equivalent to one based on reaction characteristics, and second as related to each individual target nucleus. This latter, somewhat unusual, basis has been selected on the grounds that it will be of greatest convenience to those working in the field who may wish to have at hand the available information for a particular target or target contaminant.

Lists of various reactions leading to particular nuclides are conveniently available elsewhere, e.g. Ajzenberg-Selove and Lauritsen (1959), Endt and Braams (1957), and the Nuclear Data Card Series as published by the National Bureau of Standards (U.S.A.).

In this review of ^{3}He induced reactions we shall omit detailed discussion of the 'fundamental' interactions of ^{3}He ions with individual nucleons, of radiative capture leading to ^{3}He and $^3\text{He}+\gamma$, as well as of elastic and inelastic scattering of ^{3}He ions on nuclides.

Similarly, high energy ^{3}He studies will be treated only briefly although references to the available literature will be provided.

Before concluding this review we present a short section where, by suggesting a number of remaining problems relevant to the study of ^{3}He induced reactions, we effectively summarize the present status of this field.

§2. Reaction Mechanisms

2.1. *Introduction*

The rather remarkable success of the Butler theory (1951, 1957) of deuteron induced reactions in fitting experimental data, even at low incident deuteron energies, suggested that ^{3}He reactions might be amenable to similar analysis. This would be expected to be particularly true for reactions involving the transfer of a single nucleon, e.g. (^{3}He, d) and (^{3}He, α) reactions, analogous to the well known (d, n) and (d, t) or (p, d) reactions respectively. More complex reactions such as (^{3}He, p) (^{3}He, n) or (^{3}He, t) involving transfer of two nucleons were expected to proceed characteristically via compound system interactions with direct mechanisms playing at most a subsidiary role.

Experimentally it has been demonstrated that only in certain isolated instances has it been possible to apply either of these simple approximations alone. The

availability of the data on ^{3}He reaction angular distributions has led to detailed theoretical studies of further refinements to the simple stripping theories. In this section the predictions of a number of these theories will be tested by comparison with the available experimental data.

2.2. *Compound System Mechanisms*

2.2.1. *Resonance region.*

Since the mean binding energy of the ^{3}He ion in compound systems in the light nuclei is roughly 17 MeV (figure 1) it would be expected that, in general, even at low energies, the level density and level widths involved in the compound system would be such as to preclude the possibility of studies at individual reaction resonances. As indicated in figure 1 two outstanding exceptions to this general statement occur for ^{3}He bombardment of ^{4}He and of ^{16}O leading to the compound systems ^{7}Be and ^{19}Ne respectively. As shown in figure 9 (§ 3.2) marked resonance structure occurs in the excitation curves for the decay products of the ^{19}Ne system but the resonances are not clearly resolved ; in a study of the ^{3}He on ^{4}He reaction at low incident energies direct radiative capture is observed (figure 36, § 4.5) and resonance phenomena corresponding to the excited states of ^{7}Be are found only at higher bombarding energies. In the very light nuclides, individual resonances have been observed despite high binding energies because of the paucity of levels in the compound systems involved. In particular detailed studies have been carried out on the ^{2}H(^{3}He, p)^{4}He and ^{2}H(^{3}H, n)^{4}He reactions in the vicinity of the 16 MeV levels in the compound nuclei ^{5}Li and ^{5}He respectively.

It has been found possible to obtain a reasonable fit to the experimental data for the above reactions using the standard resonance formalism (Vogt 1958, Devons and Goldfarb 1957). The resonance parameters which are obtained are listed in table 3. It is improbable that any similar clearly defined resonance structure amenable to analysis exists in ^{3}He reactions other than these rather unique cases because of the high compound nucleus excitation generally achieved even with low ^{3}He energies. The very large values of the reduced ^{3}He widths in table 3 is further evidence that we are here dealing with very special cases. In general one expects ^{3}He induced reaction widths to be small compared with those for α-particles or nucleons because the low binding energy of 2·57 MeV per nucleon (Mattauch *et al.* 1956 a) of ^{3}He makes it an unlikely unit to be found inside a nucleus. Experimental information on the average values of ^{3}He reduced widths would be of interest but such data are difficult to obtain in the absence of clearly defined resonance structure and because ^{3}He reactions in part proceed via direct interaction.

The details of the resonance fitting procedure used to obtain the parameters listed in table 3 will not be given here but may be found by referring to the original papers (see also § 4.2 for references). In order to define the parameters given in the table it is useful, however, to quote a simplified one-level Breit–Wigner relation ; for a more general discussion of resonance reactions see Vogt (1958) and Gove (1958). The cross section for a reaction involving incident ^{3}He ions and outgoing particles y may be written as :

$$\sigma = \frac{\pi}{k^2}\frac{2J+1}{2(2I+1)}\frac{\Gamma_{^3\mathrm{He}}\Gamma_y}{(E-E_0-\Delta_l)^2+\frac{1}{4}\Gamma^2} \qquad \ldots\ldots(1)$$

Table 3. ^{3}He Induced Reaction Resonance Parameters

θ_x^2 is defined as the ratio of the reduced partial reaction width to the Wigner limit, $\gamma_w^2 = \hbar^2/\mu R_0^2$, Γ_w is the most probable gamma radiation de-excitation width as deduced by Wilkinson from a systematic analysis of the experimental data.

Reaction	E_{Res} (MeV)	Γ_{Res} (MeV)	σ_{Res} (mbn)	E_{ex} (MeV)	J	$\gamma^2_{^3He}$ (kev)	$\theta^2_{^3He}$	$\gamma^2_{^4He}$ (kev)	$\theta^2_{^4He}$	Γ_γ (ev)	Γ_γ/Γ_w (e)	R (fermi)	Ref.
^{1}H (^{3}He, p) ^{4}He	0·640		695	16·81	$3/2^+$	2930	2·5	42	0·03	11	0·34	5	(*a*) (*b*)
^{4}He (^{3}He, ^{3}He) ^{4}He	5·17	0·180		4·53	$7/2^-$	680	0·36					4·4	(*c*)
^{16}O (^{3}He, ^{4}He) ^{15}O	2·400	0·355	26	10·436	$1/2^{(+)}$	281	0·67	15	0·05	0·7	0·04	5·7	(*d*)
^{16}O (^{3}He, ^{4}He)^{15}O	2·425	0·045	7	10·457	$5/2^{(+)}$	11	0·03	3·5	0·01				(*e*)

References : (*a*) W. E. Kunz (1955), (*b*) J. M. Blair, N. M. Hintz and D. M. Van Patter (1954), (*c*) P. D. Miller and G. C. Phillips (1958), (*d*) D. A. Bromley, J. A. Kuehner and E. Almqvist (1959 b), (*e*) D. H. Wilkinson (1958).

where k is the wave number of the incident ^{3}He, J and I are the spins of the compound and target nucleus respectively, E is the bombarding energy, E_0 is the formal resonance energy, Δ_l is the Thomas level shift (Lane and Thomas 1958), Γ is the total width of the compound state, $\Gamma_{^3He}$ and Γ_γ are the ^{3}He width and particle y width respectively. The particle widths $\Gamma_{^3He}$ and Γ_γ are related to the corresponding reduced widths $\gamma_{^3He}{}^2$ and $\gamma_\gamma{}^2$ through the penetration factor P_c (Lane and Thomas 1958, Vogt 1958)

$$\Gamma_\gamma = 2P_c\gamma_\gamma{}^2. \qquad \ldots\ldots(2)$$

Because of the large widths and low incident energies involved in the observed ^{3}He resonance reactions the usual approximations of assuming that the level shift Δ_l and the penetration factor P_c are constants (i.e. independent of energy over the

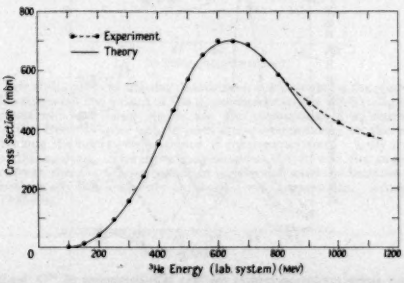


Figure 8. Excitation data for the ^{3}H (^{3}He, p) ^{4}He reaction. The solid points are the experimental data; the full curve is the fitted curve obtained using the standard resonance formalism and the parameters listed in table 3. The fact that the experimental data points lie above the single level curve at higher energies is attributable to the effects of higher resonances. From Kunz (1953, 1955).

resonance) cannot be made. The formalism used to obtain the parameters given in table 3 therefore explicitly included the energy dependence of Δ_l and P_c (Lane and Thomas 1958).

The resonance formalism has been applied by a number of authors (see table 10, §4.2 for detailed references) to the ^{3}H (^{3}He, p) ^{4}He data; for incident centre-of-mass energies less than 500 kev the single level formulation provides an adequate representation of the data with the parameters listed in table 3. At higher energies the predicted cross sections are significantly smaller than those observed consistent with the appearance of contributions from other levels in the compound system. Figure 8 shows the excitation function for this reaction as measured by Kunz (1953, 1955). The observation of the radiative capture process ^{3}He (d, γ) ^{5}Li at the same resonance with a cross section of 0·05 mbn corroborates the surprisingly large value of the ^{3}He width (Blair, Hintz and Van Patter 1954, Kunz, Butler and Holmgren 1955 a, 1955 b). Further confirmation is provided by the studies of Freier and Holmgren (1954) of the elastic scattering of deuterons and ^{3}He in the vicinity of the resonance. They find strong destructive interference between Coulomb and

resonant scattering which reduces the total elastic cross section by 40% from the full Coulomb value.

Alternate, and somewhat more general, attempts at fitting these excitation curves have been carried out by Hale and Present (1955). In effect these authors use equation (1) rewritten explicitly in terms of the real and imaginary parts of the logarithmic derivative of the nuclear wave function on the interaction surface. While they succeed in obtaining a good fit to the excitation curve over a wider energy range $E_{^3\mathrm{He}} \lesssim 1$ MeV, as compared to the single level formula, they are unable to obtain reliable results for the functional dependence of the logarithmic derivative on the incident energy.

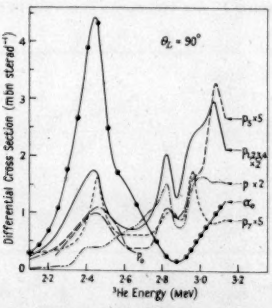


Figure 9. Proton and alpha excitation curves for ^{3}He bombardment of ^{16}O, leading to the low levels of ^{18}F and to the ground state of ^{18}O. From Bromley, Kuehner and Almqvist (1959 b).

Miller and Phillips (1958) have carried out a detailed phase shift analysis of the ^{4}He(^{3}He, ^{3}He)^{4}He elastic scattering data at several angles of observation to arrive at the resonance parameters listed in table 3.

Bromley, Kuehner and Almqvist (1958 a, b ; 1959 b) have used the detailed two-level resonance formalism, including the energy variation of the resonance parameters, to obtain fits to the coefficients in Legendre polynomial expansions to the alpha particle angular distributions for the ^{16}O(^{3}He, α)^{15}O reaction measured as a function of ^{3}He energy over the region of the two interfering resonances shown in figure 9. This is a particularly favourable case for the two-level analysis since the restriction to channel spin $^1/_2$ in both entrance and exit channels introduces marked simplifications in the formalism (Devons and Goldfarb 1957). Figures 10 and 11 show the resultant fits obtained to the experimental data using the parameters listed in table 3. The large negative P_2 term in the distributions results entirely from interference between the resonance amplitudes ; it has been necessary to postulate the dashed ramp shown in figure 11 to correspond to the effects of tails of higher resonances as shown in the excitation curves for this reaction given in figure 9.

It should be emphasized that the resonance phenomena just considered are anomalous in ^{3}He induced reactions. However, they do provide evidence

^{3}He widths can be large and that compound nucleus formation may be important in ^{3}He reactions. Other cases where less clearly defined resonance structures are observed are discussed in the following sections.

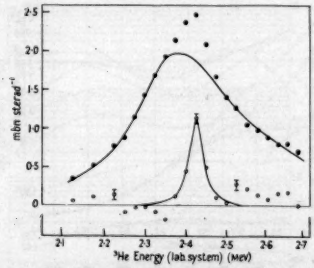


Figure 10. Ground state alpha particle angular distribution coefficients for the reaction ^{16}O (^{3}He, α) ^{15}O. The solid circles represent the values of the a_0 coefficient in the $W(\theta) = \Sigma_n a_{2n} P_{2n}(\theta)$ Legendre polynomial expansion; the open circles are the corresponding a_4 coefficient values. The full curves are the Breit–Wigner single level fits corresponding to the parameters listed in table 3 and including the energy dependence of these parameters. Only the $J = 1/2$ resonance fit is shown for the a_0 data. The difference between this fit and the experimental data is the a_0 contribution from the $J = 5/2$ resonance in agreement with the location and width of the a_4 curve which theoretically can arise only from the $J = 5/2$ resonance. From Bromley, Kuehner and Almqvist (1959 b).

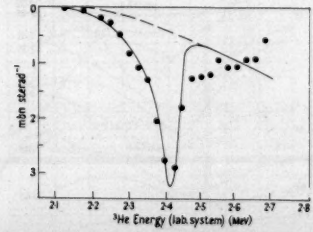


Figure 11. Coefficient of the second order Legendre polynomial expansion to the ground state alpha particles from the ^{16}O (^{3}He, α) ^{15}O reaction. The solid curve was obtained by substituting the resonance widths and locations as determined from figure 10 into the two-level formalism, including, empirically, the dashed ramp to include interference effects of higher energy resonances in this reaction. From Bromley, Kuehner and Almqvist (1959 b).

2.2.2. *Intermediate region.*

The occurrence of significant structure in the excitation curves for ^{3}He induced reactions in ^{6}Li, ^{10}B, ^{12}C and ^{13}C for incident energies $\lesssim 6$ MeV (see § 3.2) indicates that these reactions proceed, in part at least, through a compound system interaction. Detailed measurements on the angular distributions of protons from ^{3}He reactions on these target nuclides, however, have been inconsistent with the assumption of either dominant compound or direct interaction mechanisms although these angular distributions showed features characteristic of both.

Figures 12 and 13 present typical angular distributions for protons from the $^{13}C(^3He, p)^{14}N$ reaction leading to formation of the $1+$, $T = 0$ ground state and 3·95 MeV state in ^{14}N respectively. It should be noted that these data are presented

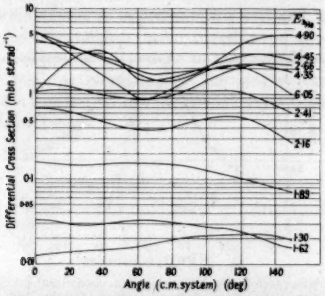


Figure 12. Proton angular distributions in the ^{13}C (^3He, p) ^{14}N ground state reaction. Note that the ordinate scale is logarithmic. Data measured at $E_{^3He} = 1{\cdot}30$, 1·62, 1·89, 2·16, 2·41 and 2·66 MeV have been reported by Bromley *et al.* (1957); those at $E_{^3He} = 4{\cdot}35$, 4·45 and 4·90 MeV by Johnston *et al.* (1958) and those at 6·05 MeV by Sweetman (1958).

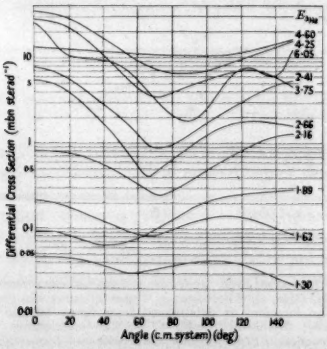


Figure 13. Proton angular distributions in the ^{13}C (^3He, p) ^{14}N reaction leading to the 3·95 MeV state in ^{14}N. Data at $E_{^3He} = 1{\cdot}30$, 1·62, 1·89, 2·16, 2·41 and 2·66 MeV have been reported by Bromley *et al.* (1957); those at $E_{^3He} = 3{\cdot}75$, 4·20 and 4·60 MeV by Johnston *et al.* (1958) and those at 6·05 MeV by Sweetman (1958).

on a logarithmic ordinate and that pronounced structure is observed particularly at higher energies. The rapid change of these angular distributions with energy suggests a compound system mechanism, as does the fact that distributions to states of identical spin and parity are quite different. Evidence for direct interaction mechanisms in this and other data will be adduced later in this section.

2.2.3. *Non-resonant region.*

In all but isolated cases, at low bombarding energies, the Q values of ^{3}He induced reactions are large enough to preclude the observation of resonant structure

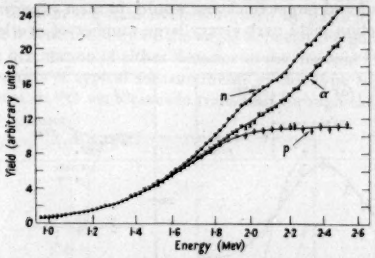


Figure 14. Excitation functions for protons ($\theta_p = 90°$), alpha particles ($\theta_\alpha = 90°$) and neutrons ($\theta_n = 0°$) from ^{3}He bombardment of ^{9}Be. The protons are those to the ground state of ^{11}B, the alpha particles are those above 5·1 Mev in energy from the ^{9}Be (^{3}He, α) ^{8}Be reaction and the neutrons from the ^{9}Be (^{3}He, n) ^{8}Be reaction were detected using a BF_3 long counter. From Almqvist *et al.* (1956).

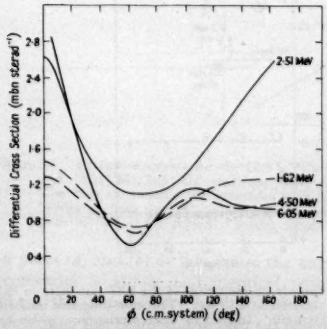


Figure 15. Proton angular distributions in the ^{9}Be (^{3}He, p) ^{11}B ground state reaction. Data at 1·62 and 2·51 Mev have been reported by Almqvist *et al.* (1956), those at 4·50 Mev by Wolicki *et al.* (1959 a) and those at 6·05 Mev by Sweetman (1958).

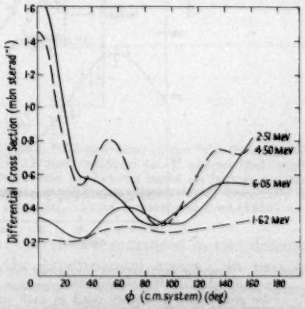


Figure 16. Proton angular distributions in the ^{9}Be (^{3}He, p) ^{11}B reaction to the 2·14 Mev state in ^{11}B. Data at $E_{^3He} = 1·62$ and 2·51 Mev have been reported by Almqvist *et al.* (1956), those at $E_{^3He} = 4·50$ Mev by Wolicki *et al.*(1959 a) and those at 6·05 Mev by Sweetman (1958).

in the excitation curves. The results shown in figure 14 for a ^{9}Be target are characteristic of these reactions. Figures 15 and 16 show angular distributions of the proton groups to the ground and first excited states of ^{11}B formed in the ^{9}Be (^{3}He, p) ^{11}B reaction. Here again the relatively rapid change with bombarding energy suggests a compound system mechanism while the persistence of structural

features such as the forward maxima, which become more pronounced with increasing energy, the 60° maximum in figure 15 and the 70° minimum in figure 16 all suggest direct interaction.

At high excitations in the compound systems it would be anticipated that the beam energy spread and target thickness would, in most experiments, be sufficient to average the reaction yield over a very large number of overlapping resonances.

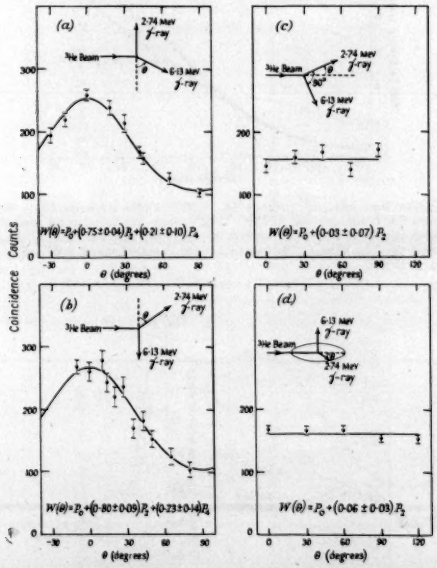


Figure 17. Angular correlations measured for the cascade de-excitation of the 8·87 MeV state in ^{14}O populated in the ^{14}N (^3He, p) ^{16}O reaction via the state at 6·13 MeV. The inset diagrams indicate the correlation geometry used in each measurement. In part (*a*) the correlation is given by $W(\theta) = P_0 + (0{\cdot}75 \pm 0{\cdot}04)\, P_2 + (0{\cdot}21 \pm 0{\cdot}10)\, P_4$; in part (*b*) by $W(\theta) = P_0 + (0{\cdot}80 \pm 0{\cdot}09)\, P_2 + (0{\cdot}23 \pm 0{\cdot}14)\, P_4$; in part (*c*) by $W(\theta) = P_0 + (0{\cdot}03 \pm 0{\cdot}07)\, P_2$ and in part (*d*) by $W(\theta) = P_0 + (0{\cdot}06 \pm 0{\cdot}03)\, P_2$. These expansions have been obtained by a least squares fitting to the experimental data. From Kuehner *et al.* (1959).

This situation has been illustrated in a detailed study of the ^{14}N (^3He, p) ^{16}O reaction (Bromley *et al.* 1957 a, 1959 e; Kuehner *et al.* 1959). In these measurements the isotropy of all the proton and gamma radiation angular distributions as well as of the proton–gamma angular correlations together with the observed lack of structure in the excitation curves suggested a very thorough averaging of the compound system level characteristics. On this hypothesis, the magnetic substates of the levels formed by particle decay of the compound levels are normally populated and

the subsequent correlations among the electromagnetic de-excitations of these levels is identical in all respects to that following their formation in radioactive decay. An experimental test of this hypothesis was adduced by noting that it required that the angular correlation between members of a double gamma-ray cascade de-exciting a level in the final nucleus formed by particle decay of the compound system to be a function only of the angle between the gamma detectors and independent of the orientation of either detector to the incident ^{3}He beam. Figure 17 illustrates the results of typical measurements of this type for the de-excitation of the 8·87 MeV level in ^{16}O via a cascade transition through the state at 6·13 MeV.

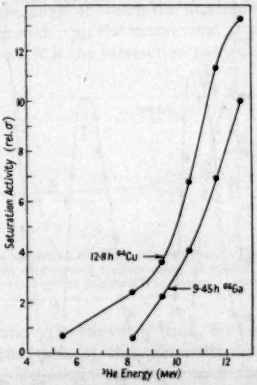


Figure 18. Excitation curves for the 12·8 h ^{64}Cu and 9·45 h ^{66}Ga produced in the ^{65}Cu (^{3}He, α) ^{64}Cu and ^{65}Cu (^{3}He, 2n) ^{66}Ga reactions with $E_{^3\mathrm{He}} = 13$ MeV incident on a copper foil stack. The ^{66}Ga excitation curve coincides with the theoretical predictions based on barrier penetrabilities as does the ^{64}Cu curve for $E_{^3\mathrm{He}} \gtrsim 9{\cdot}0$ MeV. At 5·4 MeV, for example, essentially no penetrability of the potential barrier would be expected. From Kundu and Pool (1953).

In parts (*a*) and (*b*) of this figure the angular correlation measured in two detector geometries is strong and identical within the experimental errors; in part (*c*) where the counters are held at a fixed separation of 90° and moved in a plane containing the beam and in part (*d*) where this plane does not contain the beam the correlations are isotropic as expected. Some of the consequences of such situations in terms of nuclear spectroscopic measurements will be discussed in § 3.2.

2.3. *Direct Interaction Mechanisms*

2.3.1. *Single nucleon transfer.*

The heading of this sub-section is of course a misnomer in the exact sense, nevertheless it serves as a convenient collective representation for the (^{3}He, α), (^{3}He, d), (^{3}He, 2p) and (^{3}He, np) reactions. These reactions have received surprisingly little experimental study because of experimental difficulties involved; in the first two cases the difficulty is that of resolving the alpha particles and

deuterons from the background of particles of lower specific ionization, and in the latter two cases coincidence measurements are required to detect the three-body break-up in unambiguous fashion.

As noted previously, the (^{3}He, α) and (^{3}He, d) reactions would be expected *a priori* to show marked resemblance to the well known (d, t) or (p, d) and (d, n) reactions in proceeding predominantly, even at low incident energies, via a direct interaction mechanism.

One of the first pieces of experimental evidence supporting this expectation was the measurement (Kundu and Pool 1953) of the excitation curves for the (^{3}He, α)

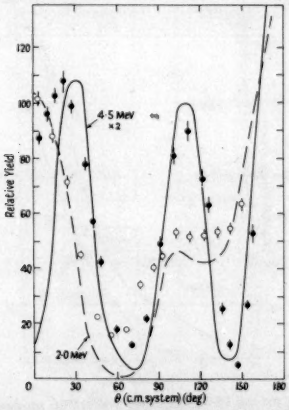


Figure 19. Alpha particle angular distributions for the ^{13}C (^{3}He, α) ^{13}C reaction. The solid points correspond to measurements at an incident energy of 2·00 MeV (Holmgren 1956, 1957); the open points to measurements at 4·50 MeV (Holmgren *et al.* 1956 c, 1957). The full and broken curves are the corresponding theoretical fits obtained by Owen, Madansky and Edwards (1959).

and (^{3}He, 2n) reactions on ^{65}Cu in the range $4 \leqslant E_{^3\mathrm{He}} \leqslant 12$ MeV (figure 18) using an activation technique with stacked foils. It was found that the (^{3}He, 2n) data fitted the compound system predictions (Blatt and Weisskopf 1952) while the (^{3}He, α) cross section had an anomalously large cross section particularly at the lower incident energies suggestive of a direct interaction contribution. Further supporting evidence is obtained from detailed measurement (Bromley *et al.* 1957) of the relative total cross sections for the (^{3}He, p), (^{3}He, n) and (^{3}He, α) reactions on ^{13}C. At incident energies in the vicinity of 2·0 MeV, the total cross sections for these reactions to the ground states of ^{14}N, ^{14}O and ^{11}C are 3·5, 5·0 and 35 millibarns steradian^{-1} respectively. The factor of 7 by which the (^{3}He, α) reaction cross section, with only 2 MeV alpha particles, exceeds the s-wave neutron emitting cross section of the (^{3}He, n) reaction strongly suggests a direct interaction mechanism for the former, with consequent reduction of the effectiveness of the Coulomb barrier in inhibiting the alpha reaction.

2.3.1.1. (^{3}He, α) *reactions.* Early measurements on the angular distributions of alpha particles from the reaction ^{13}C(^{3}He, α)^{12}C at 2·0 Mev (Holmgren 1956, 1957) and 4·5 Mev (Holmgren *et al.* 1956 c, 1957) showed very marked structure in these distributions as shown in figure 19. While these distributions do show the characteristic forward peaking expected for normal Butler stripping it is impossible to obtain any measure of agreement with the data using this formalism which involves transfer of a single neutron, alone. That this is not surprising follows immediately from consideration of simple conservation of linear momentum in the reaction (Huby 1952). As illustrated in figure 20, for (^{3}He, α) reactions with low Q values (i.e. low alpha particle energies) the vector diagram closes readily, thereby defining semi-classically the angle at which the maximum in the stripping angular distribution might be expected. p_n, the momentum of the transferred neutron, is given by $l_n h = R \times p_n$ where R is the interaction radius, and l_n is the characteristic

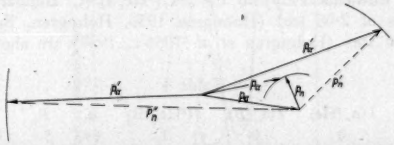


Figure 20. Huby momentum diagrams for (^{3}He, α) reactions. The unprimed, single primed and double primed situations correspond to increasing Q values, hence increasing energy of the emergent alpha particle for a given incident ^{3}He energy. From Huby (1952).

orbital neutron momentum. On the other hand, for large Q value reactions, as is usual for incident ^{3}He ions (table 2), this vector diagram does not close except for very high values of p_n. To obtain closure at back angles in the ^{13}C(^{3}He, α)^{12}C reaction, where maxima are found experimentally in the distribution, would require the pick-up of a neutron with approximately 120 Mev energy from the target nucleus. The very low probability of finding such a wave function component is reflected in a vanishing amplitude for the stripping process.

Two alternate mechanisms have been proposed to explain the existence of backward maxima in direct interactions. In recent semi-classical theoretical treatments of the direct interaction mechanism (McCarthy 1959, Butler, Austern and Pearson 1958, Austern 1959) it has been pointed out that reflection from the back surface of the target nucleus can result in imaging of a forward maximum in the differential cross section into a backward maximum. Reactions in the nuclear interior which correspond to forward particle emission at an angle θ to the incident beam direction, as a result of the surface reflection, also correspond to particle emission in a cone about the direction $\pi - \theta$, where the cone half angle is given by the critical angle of reflection $\phi_c \sim 30°$ for emerging nucleons for example. Because of the smearing inherent in the finite angle of the cone this process would not be expected to reproduce any structure in the forward maximum, consequently this process appears inadequate to explain other than a general increase in the differential cross section with increasing angle, and then only when accompanied by strong forward peaking.

In fitting the experimental angular distributions for the $^{11}B(d, n)^{12}C$ reaction which also show large backward maxima, Owen and Madansky (1957) introduced the concept of 'heavy particle' or 'backward' stripping wherein the target is considered as a complex consisting of a core and the emergent particle in the reaction. The incident particle interacts and fuses with the core while the emergent particle plays a spectator role and in consequence is preferentially emitted in a backward direction. In general, both normal stripping and heavy particle stripping amplitudes contribute to the total reaction yield and thus interfere. Since the emergent particle comes from either the incident or target nucleus it becomes essential to use properly antisymmetrized wave functions for the system as contrasted to the case of simple stripping alone.

The reader is referred to the series of papers by Owen and Madansky (1957, 1957 a, 1958, Owen, Madansky and Edwards 1959) and Fulton and Owen (1957) for details of the formalism required in these analyses. The fits obtained (Owen, Madansky and Edwards 1959) to the $^{13}C(^3He, \alpha)^{12}C$ angular distributions at incident energies of 2·00 MeV (Holmgren 1956, Holmgren, Bullock and Kunz 1956 a) and 4·50 MeV (Holmgren *et al.* 1956 c, 1957) are shown in figure 19.

Table 4

E (^{3}He)	l (n, ^{12}C)	l (n, ^{3}He)	l (α, ^{9}B)	l (^{3}He, ^{9}B)	a_1	R_1	a_2	R_2	A_2/A_1
2·0 MeV	1	0	0	1	4·6	5	4·3	5·5	0·50
4·5 MeV	1	0	0	1	4·2	4·8	4·4	7·2	1·78

Stripping parameters corresponding to the fitted curves for the ^{13}C (^3He, α) ^{12}C reaction shown in figure 19.

The parameters used in these fits are listed in table 4. As indicated, in both cases, a p wave neutron is stripped from the ^{12}C core in the forward process with the neutron and ^3He in an S state appropriate to the formation of the emergent alpha particle. In the heavy particle stripping process it has been assumed that ^{13}C can be represented by the core nuclide ^9B and ^4He in an S state and that the ^3He incident ion is captured into a P state by the ^9B core. R_j and a_j are the cut-off radii used in the integrations over the momentum functions in the stripping and heavy particle stripping processes respectively (Butler 1957); it is gratifying to find that the values required for these parameters are quite comparable to the interaction radius in the problem. A_1 and A_2 are the amplitudes of the stripping and heavy particle stripping processes respectively; the increase of the A_2/A_1 ratio with the incident energy is consistent with the requirement for capture of a charged ^3He ion in the heavy particle stripping case and only a neutron transfer in the normal stripping case. It should be borne in mind that these calculations have ignored all nuclear and Coulomb distortion of the assumed plane wave representation both of the incident and outgoing particles. At the low energies involved it might have been expected *a priori* that such distortion effects would be very important. Interference between the ordinary and heavy particle stripping amplitudes plays a major role in determining the fit to the experimental data.

It has been suggested by Satchler (1959, private communication) that detailed measurements of the (^{3}He, p), (^{3}He, d), (^{3}He, ^{3}He) and (^{3}He, α) reactions on any given target nucleus leading to measurements of the relative stripping and heavy

particle stripping amplitudes provide crucial information on the existence of nucleon clusters on the nuclear surface. Such clustering may be important not only for the wave functions but also for the total energies of the system.

Owen and Madansky (1958) have noted that since the recoil axes for the residual nucleus do not coincide for cases of forward and backward stripping, a study of the (^{3}He, $\alpha\gamma$) correlation, which on a direct interaction picture has the recoil as a symmetry axis, would provide yet another measure of the relative amplitudes involved. Measurements on the ^{13}C (^{3}He, $\alpha\gamma$) ^{12}C* reaction (Holmgren, Wolicki and Johnston 1959) have not been amenable to such simple analysis. Correlations

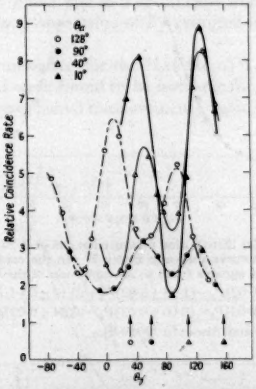


Figure 21. Angular correlations measured for the ^{13}C (^{3}He,α) ^{13}C* (γ)^{12}C reaction via the 4·43 MeV excited state of ^{12}C. The curves in the figure have been measured for different angles of alpha particle detection. The vertical arrows show the calculated angles of nuclear recoil ; in a plane wave approximation, direct interaction calculation these angles are predicted to correspond to symmetry axes for the correlations. Clearly this is not the case. Note the approximate symmetry relative to $\theta_\gamma = 90°$ for the correlations measured with alpha particles detected at $\theta_\alpha \lesssim 90°$. From Holmgren, Wolicki and Johnston (1959).

measured at $E_{^3He} = 4{\cdot}5$ MeV with the alpha particles detected at angles $\leqslant 90°$ are qualitatively similar in that they are roughly symmetric with respect to the normal to the incident beam and in phase ; for $\theta_\alpha > 90°$ there is an apparent phase shift of about 180°. The correlation yield varies by a factor of 5 in going from $\theta_\gamma = 90°$ to $\theta_\gamma = 130°$, for example, when $\theta_\alpha = 10°$. There is no evidence for symmetry relative to the individual stripping recoil axes. Typical correlations are shown in figure 21.

These data have not been subjected to a full analysis as yet in any attempt to disentangle components resulting from ordinary and heavy particle stripping. Any such analysis will be relatively difficult since the correlation measurements are much more sensitive to interference and distorted wave effects than are the corresponding angular distributions (Austern 1959). In the particular case of the correlation corresponding to $\theta_\alpha = 0°$, as discussed in § 3.3, the correlation is predicted

to be independent of the reaction mechanism involved. The structural features of the measured correlation for $\theta_\alpha = 10°$ are in accord with the theoretical predictions.

As discussed in the previous sub-section, in contrast to the (^{3}He, α) reaction on ^{13}C, that on ^{16}O at incident energies of about 2·5 MeV (Bromley, Kuehner and

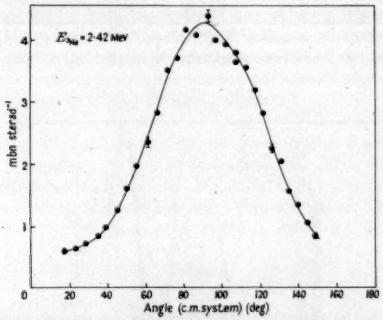


Figure 22. Alpha particle angular distribution measured at the peak of the resonance structure in the ^{16}O (^{3}He, α) ^{15}O excitation curve shown in figure 9. In the centre-of-mass system the angular distribution has been least squares fitted to an expansion of the form

$$W(\theta) = \sigma_0\{1-(0{\cdot}06\pm0{\cdot}02)P_1-(1{\cdot}13\pm0{\cdot}02)P_2+(0{\cdot}12\pm0{\cdot}03)P_3+(0{\cdot}32\pm0{\cdot}03)P_4 + (0{\cdot}02\pm0{\cdot}03)P_5-(0{\cdot}00\pm0{\cdot}03)P_6-(0{\cdot}04\pm0{\cdot}03)P_7-(0{\cdot}02\pm0{\cdot}03)P_8\}.$$

From Bromley, Kuehner and Almqvist (1959 b).

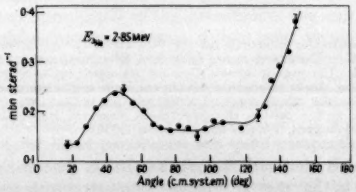


Figure 23. Alpha particle angular distribution measured at the minimum of the excitation curve for the ^{16}O (^{3}He, α) ^{15}O ground state reaction shown in figure 9. These data have been least squares fitted to the expansion

$$W(\theta) = \sigma_0\{1-(0{\cdot}37\pm0{\cdot}03)P_1+(0{\cdot}52\pm0{\cdot}05)P_2-(0{\cdot}53\pm0{\cdot}05)P_3+(0{\cdot}02\pm0{\cdot}05)P_4 -(0{\cdot}03\pm0{\cdot}05)P_5-(0{\cdot}01\pm0{\cdot}05)P_6+(0{\cdot}11\pm0{\cdot}05)P_7+(0{\cdot}01\pm0{\cdot}05)P_8\}.$$

From Bromley, Kuehner and Almqvist (1959 b).

Almqvist 1958 a, b, 1959 b) is entirely consistent with a compound system mechanism. Angular distributions measured on the resonance structure of figure 9 show characteristic symmetry about 90° in the centre-of-mass system ; figure 22 shows a typical angular distribution. At slightly higher energy, at the minimum in the alpha particle excitation curve, the angular distribution of figure 23 was

measured. The lack of symmetry and the increase in the cross section at backward angles are qualitatively similar to those observed in the $^{13}C(^3He, \alpha)^{12}C$ reaction data of figure 19, suggesting that the $^{16}O(^3He, \alpha)^{15}O$ reaction has a small direct interaction amplitude, relatively independent of energy in the region studied, which is almost completely masked at strong compound system resonances. Support for this hypothesis follows from the fact that in obtaining a fit to the alpha angular distribution data it was necessary to modify the potential phase shift (Vogt 1958) to :

$$\phi = \phi_l^0 + \tan^{-1}(R^\infty P_l / 1 - R^\infty \Delta_l) \qquad \ldots\ldots(3)$$

where $\phi_l^0 = \tan^{-1}(F_l/G_l)$ is the usual undistorted value. In this reaction the added term is a ' distorted wave ' effect due to a specifically nuclear potential rather than the Coulomb field. The non-zero value of R^∞ required is characteristic of a direct interaction mechanism.

Angular distributions measured for the $^7Li(^3He, \alpha)^6Li$ alpha particle groups to the first three states of ^6Li were found to be isotropic for $E_{^3He} = 2{\cdot}1$ MeV (Bromley, Kuehner and Almqvist unpublished measurement), again suggestive of a compound system mechanism.

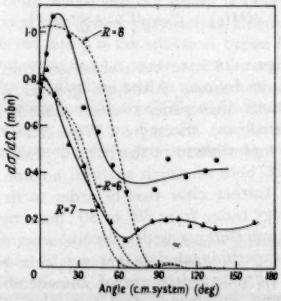


Figure 24. Deuteron angular distributions measured for the $^{11}B(^3He, d)^{12}C$ reaction at $E_{^3He} = 4{\cdot}50$ MeV. The triangles represent data measured for the deuteron group to the ^{12}C ground state and the circles for that to the 4·43 MeV excited state. The broken curves are those predicted by the Newns formalism using the cut off radii (in fermis) indicated on the figure. From Holmgren, Wolicki and Johnston (1959).

2.3.1.2. (^{3}He, d) *reactions.* Holmgren, Wolicki and Johnston (1957, 1959 a) have measured the angular distribution of deuterons from the $^{11}B(^3He, d)^{12}C$ reaction to the ground and first excited states at 4·5, 5·4 and 5·5 MeV respectively. These are shown in figure 24. In contrast to the (^{3}He, α) results just discussed these data suggest almost pure forward stripping ; the excitation curves in this energy region show no structure, consistent with such an explanation.

The full curves in these figures are drawn through the experimental points ; the broken curves were obtained from the formalism given by Newns (1952). This is a Born approximation calculation assuming a well defined interaction radius and neglecting all distortion of the assumed incident and outgoing plane waves by either nuclear or Coulomb effects. Throughout this section such situations

are referred to as stripping, or 'ordinary' stripping. The differential cross section is predicted to be of the form :

$$\sigma(\theta) = M \exp\left(\frac{-q^2}{4\gamma^2}\right) \times \left| \frac{j_l(\phi r_p)}{r_p} - \left[\frac{j_l(\phi r_p)}{h_l(k_p r_p)}\right] \frac{h_l(k_p r_p)}{r_p} \right|^2_R \quad \ldots\ldots(4)$$

where

$$\mathbf{q} = \mathbf{k}_d - \tfrac{2}{3}\mathbf{k}_{^3He} \quad \ldots\ldots(5)$$

$$\boldsymbol{\phi} = \mathbf{k}_{^3He} - \tfrac{11}{12}\mathbf{k}_d \quad \ldots\ldots(6)$$

$$k_p = i\left(\frac{2}{\hbar^2}\mu_p(Q+\epsilon)\right)^{\frac{1}{2}}. \quad \ldots\ldots(7)$$

$\mathbf{q}$ and $\boldsymbol{\phi}$ are the momentum transfer vectors for the deuteron and target respectively ; $\mathbf{k}_d$, $\mathbf{k}_{^3He}$ and $\mathbf{k}_p$ are the centre-of-mass wave numbers of the deuteron, ^{3}He ion and captured proton respectively ; Q is the energy release in the reaction and ϵ the proton binding energy in ^{3}He, μ_p is the reduced mass of the captured proton. The matrix element M is a multiplicative, angle-independent factor and a Gaussian wave function has been assumed for the ^{3}He ion :

$$\psi(^3\text{He}) = C \exp\left(-\gamma^2 \sum_{ij}{}' R_{ij}{}^2\right). \quad \ldots\ldots(8)$$

The broken curves in figure 24 have been obtained from equation (4) using the indicated reaction radii in fermis. Although general qualitative agreement is found there are very serious discrepancies which make it impossible to obtain an adequate fit with this formalism. Holmgren, Wolicki and Johnston (1959 a) find, for example, that it is not possible to fit the narrow peak at 15° without assuming either $R_0 > 15$f or $\gamma < 0{\cdot}1$f, both of which are well outside the expected range for these parameters. It is further clear that in order to fit the data the predicted cross sections in the energy range from 30° to 50° must be decreased, presumably as a result of destructive interference with a second reaction amplitude.

There are too few experimental data yet available to suggest the required modifications in detail. Further data are of considerable interest in order to elucidate the mechanisms involved ; the fact that these distributions for $E_{^3He} > 4{\cdot}5$ Mev, at least qualitatively, are susceptible to forward stripping analysis suggests that this reaction constitutes a potentially very useful tool in nuclear spectroscopy. Kuehner, Almqvist and Bromley (1958, unpublished) have found that the angular distribution of the deuteron group to the ground state of ^{11}C formed in the ^{10}B(^{3}He, d)^{11}C reaction at $E_{^3He} = 2{\cdot}1$ Mev shows strong maxima at $\sim 17°$ and 75° similar to those in the alpha angular distributions of figure 19, suggesting that both ordinary and heavy particle stripping play significant roles at this energy.

The ^{11}B(^{3}He, dγ)^{12}C* angular correlation has been studied by Wolicki *et al.* (1959) as a function of the angle of deuteron observation. As in the (^{3}He, $\alpha\gamma$) case (figure 21) these data are not amenable to any single analysis relative to the recoil directions and further more detailed measurements are required.

2.3.1.3. (^{3}He, 2p) *and* (^{3}He, np) *reactions.* Apart from the early measurements of Fremlin (1952) there are essentially no experimental data available on these three-body reactions. Demeur (1954) has obtained theoretical predictions for the

angular correlation of the reaction products detected in coincidence, assuming a direct interaction mechanism. The predicted correlation is of the form

$$W(k_1, k_2) = A\, G^2(\mathbf{K}_1, \mathbf{K}_2) \sum_{l=0}^{\infty} l\Lambda_l j_l^2(kR) \qquad \ldots\ldots(9)$$

where

$$\mathbf{k} = \mathbf{k}_{^3\mathrm{He}} - \frac{M_T}{M_R}(\mathbf{k}_1 + \mathbf{k}_2) \qquad \ldots\ldots(10)$$

is the momentum transfer to the target,

$$\mathbf{k}_j = \frac{M_j}{\hbar}.\mathbf{v}_j \qquad \ldots\ldots(11)$$

are the centre-of-mass wave numbers of the emitted nucleons,

$$\mathbf{K}_j = \mathbf{k}_j - \tfrac{1}{3}\mathbf{k}_{^3\mathrm{He}} \qquad \ldots\ldots(12)$$

are the momentum change vectors of the emitted nucleons.

$G(K_1, K_2)$ is a form factor dependent upon the detailed internal wave function assumed for the ^{3}He ion. Demeur (1954) has examined the question of selecting appropriate G values in some detail and the reader is referred to this paper. A, a multiplicative constant, is a function of the reaction kinematics, Λ_l is a function of the matrix elements involved and j_l is the spherical Bessel function of order l.

Experimental data are required on these reaction correlations, both to check on the applicability of the calculations and to give information on the ^{3}He internal wave function through determination of $G(K_1, K_2)$. It should be noted that the photographic plate technique used by Fremlin (1952) automatically provides a coincidence measurement in the (^{3}He, 2p) reaction but does not suffice in the (^{3}He, np) case.

2.3.2. *Two-nucleon transfer.*

2.3.2.1. (^{3}He, p), (^{3}He, n) *and* (^{3}He, t) *reactions—survey.* As noted previously, in the discussion of compound system interactions, angular distributions of protons and of neutrons from ^{3}He induced reactions at incident energies $\lesssim 6$ MeV show characteristics of both compound and direct interaction mechanisms. Figures 12, 13, 14, and 15 illustrate this behaviour.

Assuming forward stripping alone, the ^{12}C (^{3}He, p) ^{14}N reaction to $1+$ states in ^{14}N (Bromley *et al.* 1957) is allowed for only S and D wave capture of a neutron–proton pair in a triplet configuration. On this basis angular distributions peaked at forward angles with a subsidiary peak between 50° and 80° would be expected; furthermore, neglecting Coulomb distortion these predictions are equally applicable to the data of figures 12 and 13. Clearly this assumption is untenable.

On the other hand, in this reaction leading to 0^+ states in ^{14}N, only S wave capture of a neutron–proton pair in a singlet state is permitted by the selection rules leading to a stripping prediction of a forward maximum. Figure 25 shows the corresponding experimental results; at ^{3}He energies above about 4 MeV this forward maximum becomes very pronounced. It should be noted that the reaction kinematics here permit closure of the Huby diagram favouring forward stripping; the data of figure 25 also suggest a heavy particle stripping amplitude which appears to have a maximum value at incident energies of approximately 5 MeV. Since the

stripping amplitudes would be expected to show a monotonic dependence on the incident energy, the backward peaking may well reflect resonance phenomena which also appear in the reaction excitation curve at these energies.

Similarly, the stripping prediction for the ^{6}Li(^{3}He, p)^{8}Be ground state reaction involving S wave capture of a triplet neutron–proton pair would be expected to show pronounced forward peaking with increasing incident energy. Figure 26 presents the experimental angular distribution data obtained by Schiffer *et al.* (1956 a, b) on this reaction. The pronounced increase in the back angle cross section with increasing ^{3}He energy is consistent with a heavy particle stripping

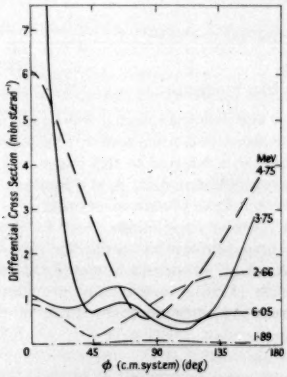


Figure 25. Proton angular distributions from the ^{13}C(^{3}He, p)^{14}N* reaction to the 2·313 Mev first excited state of ^{14}N. Measurements at $E_{^{3}\mathrm{He}} = 1{\cdot}89$ and 2·66 Mev have been reported by Bromley *et al.* (1957); those at 3·75 and 4·75 Mev by Johnston *et al.* (1958) and those at 6·05 Mev by Sweetman (1958).

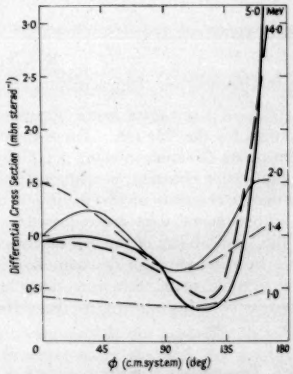


Figure 26. Proton angular distributions from the ^{6}Li(^{3}He, p)^{8}Be reaction leading to the ground state of ^{8}Be. From Schiffer *et al.* (1956 a, b).

hypothesis. The kinematics of this reaction are such that there is essentially zero probability of finding a neutron–proton pair in the ^{3}He complex with sufficient energy to permit the closure of the Huby diagram (figure 20) in agreement with the experimental observation that the stripping amplitudes are relatively small.

These figures illustrate the utility of this diagram in presenting a convenient representation of the reaction mechanism.

2.3.2.2. *Final state statistical factors.* From an analysis of the cross-sectional data on the ^{6}Li(^{3}He, p)^{8}Be and ^{10}B(^{3}He, p)^{12}C reactions in the energy range from 3 to 5 Mev, Schiffer (1956, private communication) has suggested a correlation between the cross-sectional magnitudes and the statistical factor $(2J+1)$, where J is the total angular momentum of the final state. Such a correlation would be indicative of a direct interaction mechanism and has been found, for example, in (d, p) and (d, n) reaction studies. Table 5 presents some of the relevant data.

The data on ^{6}Li(^{3}He, p)^{8}Be are taken from the work of Schiffer *et al.* (1956 a, b) ($3 \lesssim E_{^3\mathrm{He}} \leqslant 5$ MeV); that on ^{10}B from Schiffer *et al.* (1956 a, b) ($3 \leqslant E_{^3\mathrm{He}} \leqslant 5$ MeV) and Gove *et al.* (1957), Almqvist *et al.* (1960) ($E_{^3\mathrm{He}} \leqslant 2{\cdot}1$ MeV); that on ^{12}C from Bromley *et al.* (1957) ($E_{^3\mathrm{He}} = 2{\cdot}16$ MeV) and on ^{14}N from Bromley *et al.* (1957 a, 1959 e) ($E_{^3\mathrm{He}} = 2{\cdot}1$ MeV). Cross-section measurements on the ^{16}O(^{3}He, p)^{18}F reaction (Kuehner, Almqvist and Bromley 1958 a, 1960) have not been included in table 5 since the marked resonance structure in the yield curves precludes a dominant direct interaction amplitude. Similarly, the relative cross sections in the ^{13}C(^{3}He, p)^{14}N reaction show local energy variations reflecting residual resonance amplitudes.

Table 5 indicates at least a qualitative correlation of the type suggested by Schiffer (1956, private communication). Further measurements, preferably at higher energies, to avoid specific resonant effects, are required to establish this correlation beyond question.

Table 5. Statistical Factors in ^{3}He Induced Reactions

Reaction	Final state E_x (MeV)	J^π	$2J+1$	$\langle\sigma\rangle$ (mbn)	$\dfrac{\langle\sigma\rangle}{2J+1}$
^{6}Li (^{3}He, p) ^{8}Be	0	0^+	1	15	15
^{6}Li (^{3}He, p) ^{8}Be	2·9	2^+	5	60	12
^{10}B (^{3}He, p) ^{12}C	0	0^+	1	1·2	1·2
^{10}B (^{3}He, p) ^{12}C	4·43	2^+	5	11	2·2
^{10}B (^{3}He, p) ^{12}C	7·76	0^+	1	1	1·0
^{10}B (^{3}He, p) ^{12}C	9·60	$(1^-, 2^+)$	3, 5	2·5†	0·8, 0·5
^{10}B (^{3}He, p) ^{12}C	10·76			0·5†	
^{10}B (^{3}He, p) ^{12}C	11·82			1·0†	
^{10}B (^{3}He, p) ^{12}C	12·78	1^+	3	1·5†	0·5
^{10}B (^{3}He, p) ^{12}C	13·37			2·1†	
^{10}B (^{3}He, p) ^{12}C	14·03			2·4†	
^{10}B (^{3}He, p) ^{12}C	15·10	1^+	3	1·5	0·5
^{10}B (^{3}He, p) ^{12}C	16·10	2^+	5	2·5†	0·5
^{10}B (^{3}He, p) ^{12}C	16·64	2^-	5	3·0†	0·6
^{12}C (^{3}He, p) ^{14}N	0	1^+	3	5·2	1·7
^{12}C (^{3}He, p) ^{14}N	2·3	0^+	1	3·5	3·5
^{12}C (^{3}He, p) ^{14}N	3·9	1^+	3	6·1	2·0
^{14}N (^{3}He, p) ^{16}O	0	0^+	1	0·06	0·06
^{14}N (^{3}He, p) ^{16}O	8·87	2^-	5	0·33	0·07
^{14}N (^{3}He, p) ^{16}O	10·36	4^+	9	0·38	0·04

† These values have been taken from the 90° differential cross sections assuming essentially isotropic angular distributions as observed for the lower states by Schiffer *et al.* (1956 a).

2.3.2.3. *Angular distribution calculations.* At ^{3}He energies less than approximately 5 MeV the experimental data suggest that except in isolated cases such as those presented, the reactions proceed in relatively complex fashion with ordinary and heavy particle, direct interactions as well as compound system amplitudes and the mutual interference terms all of comparable magnitude and importance. As the energy is increased the compound system amplitudes for a given reaction

channel become less important both because of competition between the larger number of open channels and because the higher energies imply higher incident orbital angular momentum components with larger impact parameters, hence relative favouring of surface or direct interactions. Furthermore, since the number of compound states involved increases, the assumed randomness of the phases of these levels strongly reduces the interference terms and the corresponding structure in the angular distributions. Some of these effects have been considered theoretically by Neudachin (1959).

El Nadi and El Khishin (1959) have recently published a theoretical study of stripping reactions in which two nucleons are transferred simultaneously.† They express the differential cross section for these reactions as:

$$\frac{d\sigma}{d\Omega} \sim \Sigma |C_1 F - C_2 B|^2 \qquad \ldots\ldots(13)$$

where F and B are the ordinary and heavy particle stripping amplitudes respectively and C_j are normalized constants.

The stripping amplitude is given by

$$|F|^2 \sim \exp\left(\frac{-q^2}{4\gamma^2}\right) \left| \hat{l}_n \times \hat{l}_p \sum_{n=0}^{\infty} [I_{n+\frac{1}{2}}(3\gamma^2 R_t^2)]\,(n+\tfrac{1}{2}) \right.$$
$$\times \sum_{m=-n}^{n} (-1)^{m_n+m_p+m} \sum_{Lm L'm'} \langle l_n n 00 | L0\rangle \langle l_n n, -m_n m | L0\rangle$$
$$\left. \times \langle l_p n 00 | L'0\rangle \langle l_p n, -m_p m | L'0\rangle\, j_L(\tfrac{1}{2}\phi_p R_t) j_{L'}(\tfrac{1}{2}\phi_p R_t) \right|^2. \qquad \ldots\ldots(14)$$

Here the subscripts n and p refer to the neutron–proton pair captured from the incident ^{3}He ion. The remaining parameters are defined as in equations (5), (6) and (7), with the obvious modifications that the wave number of the outgoing deuteron k_0 in equations (5) and (6) is replaced by that of the outgoing proton (as distinct from that captured). $I_{n+\frac{1}{2}}$ is the modified Bessel function of order $n+\frac{1}{2}$ tabulated by Petiau (1955), for example. The selection rule applicable to the stripping amplitude is:

$$\mathbf{J}_{\text{final}} = \mathbf{J}_{\text{initial}} + \mathbf{l}_n + \mathbf{l}_p + \tfrac{1}{2} + \tfrac{1}{2}. \qquad \ldots\ldots(15)$$

In similar fashion, the heavy particle stripping amplitude B, where the emergent proton comes from the target rather than the incident ^{3}He ion, is given by

$$B \sim i^{l_c + l_{p'}} Y^*_{l_{p'}, m_p}(\mathbf{b}.\mathbf{B}) j_{l_{p'}}(bR_b) j_{l_c}(BR_b) \qquad \ldots\ldots(16)$$

where the subscripts c and p refer to the core c and the outgoing proton p which together comprised the initial target nuclide. The vectors $\mathbf{b}$ and $\mathbf{B}$ are the centre-of-mass momentum transfers which may be obtained directly from the vector diagram for the reaction:

$$\mathbf{b} = -\left(k_p + \frac{1}{M_t} k_{^3\text{He}}\right) = \begin{matrix}\text{momentum change of} \\ \text{emitted proton}\end{matrix} \qquad \ldots\ldots(17)$$

$$\mathbf{B} = -\left(\frac{3}{M_t} k_p + k_{^3\text{He}}\right) = \begin{matrix}\text{momentum change of} \\ \text{captured } ^3\text{He ion}\end{matrix} \qquad \ldots\ldots(18)$$

† *Note added in proof.* As pointed out by H. C. Newns, in a paper soon to be published, the results obtained by El Nadi and El Khishin (1959) are invalidated by an error in summing and averaging over the magnetic quantum numbers of initial and final states.

l_c is the orbital momentum with which the ^{3}He ion is captured by the core, l_p' is that of the proton in the target nucleus relative to the core.

Figure 27 shows the fits obtained by these authors to the $^{12}C(^3He, p)^{14}N$ angular distributions of protons to the first and second excited states of ^{14}N,

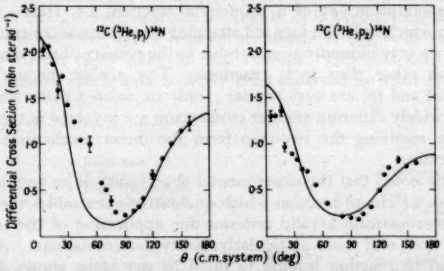


Figure 27. Theoretical fits obtained to the proton angular distributions from the ^{13}C (^3He, p) ^{14}N reaction leading to the first and second excited states of ^{14}N. These data have been obtained for $E_{^3He} = 4{\cdot}50$ MeV ; the parameters corresponding to the full curves are given in table 7. From El Nadi and El Khishin (1959).

measured for $E_{^3He} = 4{\cdot}5$ MeV. The parameters used in obtaining these fits are listed in table 6. The values of l_n and l_p follow from the selection rule of equation (15) ; l_c is determined as 1 or 3 from the data of table 7.

Table 6

Reaction	Stripping amplitude			Heavy particle stripping amplitude		
	l_p	l_n	R_f (fermis)	l_c	l_p'	R_b (fermis)
^{13}C (^3He, p_1) $^{14}N^*$ 2·31 MeV state	0	0	4·8	1	1	5·8
^{13}C (^3He, p_2) ^{14}N 3·95 MeV state	0	0	4·8	1	1	5·5

Parameters used in obtaining the theoretical curves shown in figure 27.

Table 7

Reaction	^{11}B	+	^3He	→	^{14}N
J	$\frac{3}{2}$		$\frac{1}{2}$		1
π	−		+		+

Angular momentum and parity balance in the ^{13}C (^3He, p) ^{14}N reaction.

Similarly, since ^{11}B and ^{13}C have opposite parity l_p' is either 1 or 3 and, in usual fashion, only the lowest orbital values permitted by the selection rules have been used in the analyses.

It should be noted, however, that although a reasonable fit is obtained to the experimental data, the values of $l_n = l_p = 0$ required for this fit are at variance with

shell model expectations where the captured nucleons would be expected to enter p orbits. Any final judgement on the general validity of this formalism must await further experimental test.

Komarov *et al.* (1959) have examined the (^{3}He, n) and (^{3}He, p) reactions theoretically both on the assumption (*a*) of simultaneous transfer of two nucleons and on the assumption (*b*) of a sequential reaction, i.e. (^{3}He − d − n). These calculations in which only the forward stripping amplitude is considered have been applied in only a very limited way as yet due to the paucity of relevant experimental information on other than (n, t) reactions. The predictions corresponding to assumptions (*a*) and (*b*) are very similar ; only in isolated cases where initial and final states of widely differing angular momentum are involved is there any hope of experimentally resolving the two-step from the direct mechanism from angular distribution studies.

It should be noted that the experimental observation of an angular distribution of protons from a (^{3}He, p) reaction which qualitatively resembles the simple Butler shape does not constitute a valid criterion for application of the usual stripping formalism for spin, and more particularly, parity determinations. As an example, the ^{16}O (^{3}He, p) ^{18}F reaction leading to the 1·70 Mev state, shows a characteristic $l_q = 1$ Butler shape which would correspond to negative parity for this state (Allen 1959, private communication). Measurements on the ^{19}F (p, d) ^{18}F reaction (Bennett 1958) together with those on the ^{16}O (^{3}He, pγ) ^{18}F reaction (Kuehner, Almqvist and Bromley 1958 a) however establish the J^π assignment to this state as 1^+ (Bromley, Kuehner and Almqvist 1958). The situation is similar to that encountered in early (dd′) measurements (Middleton and Tai 1952) where, for example, application of Butler formalism to the ^{20}Ne (dd′) results, suggested by observation of typical $l = 1$ angular distributions, suggested negative parity for the first excited state of ^{20}Ne.

Angular correlation measurements (Johnston, Holmgren and Wolicki 1959) on the ^{10}B (^{3}He, pγ) ^{12}C* reaction at $E_{^3\mathrm{He}} = 4{\cdot}5$ Mev, in common with all the correlation measurements available to date, are not amenable to any simple analysis.

It should be noted that the ^{3}He (d, p) ^{4}He reaction studied by Allred (1951) and the mirror reaction T (d, n) ^{4}He studied by Brolley *et al.* (1951) showed angular distributions of alpha particles which were essentially identical. This identity was interpreted by Brolley *et al.* (1951) as evidence for the equality of n–n and p–p forces, i.e. the hypothesis of charge symmetry of nuclear forces. Following the development of the Butler stripping theory it was noted by Butler and Symonds (1951) that the data on these reactions were explicable in terms of simple forward stripping formalism independent of any more detailed assumptions. No significant deviations from the forward stripping predictions were observed as would be the case if any appreciable reaction amplitude resulted from the stripping of a proton from the ^{3}He nucleus analogous to a forward stripping in a (^{3}He, p) reaction.

Because the Q values characteristically are negative for (^{3}He, t) reactions, no detailed studies of this reaction have been reported as yet at low ^{3}He energies. As will be discussed in the following subsection, measurements at higher ^{3}He energies suggest a direct nuclear surface excitation mechanism for this, and other, ^{3}He induced reactions.

2.3.3. *Direct surface excitation.*

It has recently been found by Wegner and Hall (1958 a, 1959, and private communication (1959)) using ^{3}He ions in the energy range from 20–30 MeV that the angular distributions of protons, deuterons and alpha particles from ^{3}He induced reactions show the diffraction oscillations which are now characteristic of a large variety of higher energy reaction studies. Figure 28 is a typical angular distribution measured for the ^{12}C(^{3}He, d)^{13}N reaction measured at $E_{^3\mathrm{He}} = 21$ MeV.

In analogy to the (pp′) (Gugelot and Phillips 1956) and (αα′) (Gugelot and Rickey 1956) results which have recently been analysed by Blair (1960) using the

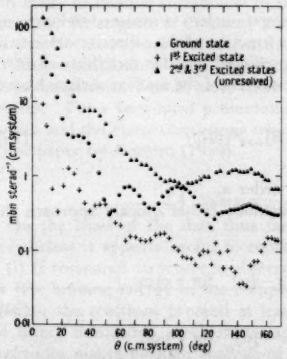


Figure 28. Angular distributions of deuterons from the ^{12}C (^{3}He, d) ^{13}N reaction at $E_{^3\mathrm{He}} =$ 21 MeV. From Wegner and Hall (1959 a).

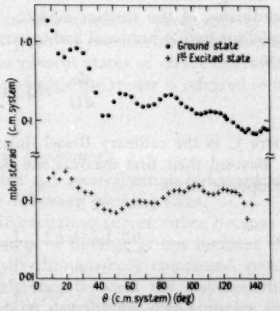


Figure 29. Angular distributions of tritons from the ^{9}Be (^{3}He, t) ^{9}B reaction at $E_{^3\mathrm{He}} = 24$ MeV. From Wegner and Hall (1959, private communication).

diffraction model of Drozdov (1955, 1958) and Inopin (1957), the characteristics of the diffraction patterns are much more closely related to the parity change in the reaction than to the exact angular momenta involved. At any given incident energy, the angular distributions of particle groups leading to states of one parity are all very similar and the diffraction oscillations are roughly 180° out of phase with those in the angular distributions of groups to states of the opposite parity.

In a study of the angular distributions of deuterons from the ^{9}Be (^{3}He, d) ^{10}B reaction at $E_{^3\mathrm{He}} = 25$ MeV for example, all angular distributions were characterized by forward peaking; groups corresponding to resolved levels showed diffraction structure at angles less than 90°. For formation of final states of even parity all distributions had maxima at 30° and 55° and minima at 40° and approximately 70° in the centre-of-mass system. No measurements are available on this reaction, on resolved groups leading to states of opposite parity; however, the angular distributions of composite groups corresponding to formation of a doublet of levels of opposite parity shows essentially no diffraction structure suggesting that the

oscillations are out of place in the individual distributions and thus cancel in the sum (Wegner and Hall 1958 a).

Similar effects are evident in figure 28 ; the angular distributions to the ground and first excited states of ^{13}N, with assignments $1/2^-$ and $1/2^+$ respectively, show the characteristic phase difference. The group corresponding to the unresolved $3/2^-$, $5/2^+$ doublet of states at about 3·5 MeV in ^{13}N shows reduced structure corresponding to the overlap of two out of phase patterns (Wegner and Hall 1958 b, 1959, and private communication (1959)). Figure 29 shows two angular distributions from the ^{9}Be(^{3}He, t)^{9}B reaction at $E_{^3\mathrm{He}} = 21$ MeV, again showing diffraction effects on this two nucleon exchange reaction (Wegner and Hall 1959, private communication).

In the Drozdov and Inopin models, the target nucleus is assumed to be strongly absorbing within a radius R which in turn is a function of the collective deformation coordinates of the nuclear surface. The predicted angular distributions in this model for both vibrational and rotational excitations of the nuclear surface have the form

$$\frac{d\sigma}{d\Omega} \sim A\,\mathrm{J}_n^2(kR\theta) + B\mathrm{J}_{n+2}^2(kR\theta) \qquad \ldots\ldots(19)$$

where J_n is the ordinary Bessel function of order n.

Beyond their first maxima the Bessel functions rather quickly approach their asymptotic forms

$$\left.\begin{aligned} \lim_{x\to\infty} \mathrm{J}_n(x) &\to \pm\sqrt{\left(\frac{2}{\pi x}\right)}\cos(x-\tfrac{1}{4}\pi) \quad \text{for } n \text{ even} \\ &\to \pm\sqrt{\left(\frac{2}{\pi x}\right)}\sin(x-\tfrac{1}{4}\pi) \quad \text{for } n \text{ odd.} \end{aligned}\right\} \qquad \ldots\ldots(20)$$

Consequently, beyond the first maximum, the squares of all odd Bessel functions have oscillations which are in phase and 180° out of phase with those of all even Bessel functions.

Since angular momentum and parity selection rules limit the values of n in equation (19) depending upon the parity of the final states involved and since the structure of the pattern is independent of the angular momenta, hence the order of the Bessel functions, this theoretical treatment provides an adequate representation of the experimental results. The validity of this diffraction model depends critically upon the mean free path of the incident ion in nuclear matter ; the shorter is this path, the more nearly is the initial assumption of a sharp nuclear surface justified. Because of the low ^{3}He internal binding energy of 7·72 MeV it would be anticipated that such a model would have particular relevance to ^{3}He reactions.

2.3.4. *Knock-on processes.*

In a photographic emulsion study of the reaction products from 30 MeV ^{3}He ion bombardment of the nuclear emulsion itself Fremlin (1953) identified reactions of the types (^{3}He, n), (^{3}He, α), (^{3}He, d), (^{3}He, t), (^{3}He, 2p), (^{3}He, pn) and (^{3}He, αp) with certainty as well as (^{3}He, 2p2α) and (^{3}He, αpn) reactions with less certainty.

The angular distributions of the charged particles in these multiple processes were predominantly forward characteristic of a direct reaction mechanism.

Ise *et al.* (1954) have used a ^{3}He ion source in the 184 in. Berkeley cyclotron to produce a beam of 330 Mev deuterons as a result of stripping 493 Mev ^{3}He ions on internal targets. External deuteron beam intensities of 5×10^{-13} amperes were obtained at this energy; the maximum deuteron energy available from direct deuteron acceleration in this cyclotron was 190 Mev.

2.3.5. *Summary—direct interaction.*

In all of the direct interaction formalisms thus far discussed, no attempt has been made to include corrections to simple Born approximation calculations. In particular, the neglect of Coulomb effects must be considered as more serious for ^{3}He than for deuteron induced reactions.

Until considerably more experimental data are available, preferably in the energy range from 5 to 20 Mev, there is little point in considering more detailed theories. For a very lucid presentation of the present status of direct interaction theories and the many corrections which are applicable the reader is referred to the review paper by Austern (1959).

2.4. *Summary—Reaction Mechanisms*

On the basis of the data thus far available any generalization is hazardous; nevertheless it appears useful to consider the following working rules.

(i) If resonance structure is observed in the excitation curves either because of low ^{3}He binding energy in the compound system or because of low incident ^{3}He energies, the reactions proceed at least partially through both compound system and direct interaction mechanisms and interference between the corresponding amplitudes makes detailed analysis of the angular distribution data complex and, in general, unprofitable since the number of unknown parameters usually exceeds the number of input data by a large margin. Exceptions of course occur when isolated resonances are found.

(ii) For incident energies in the range 5 to approximately 15 Mev the data appear to be amenable to analysis using current direct interaction formalisms including both forward and backward stripping. Much more experimental data, both on angular distributions and correlations, are needed to establish the mechanisms. Once established, reactions in this energy range will provide not only nuclear spectroscopic data such as level assignments, but also crucial information on ' cluster ' behaviour and correlations in nuclear matter.

(iii) At higher incident energies, because of the relatively short mean free path of the ^{3}He ions, the various reactions lose their specific identities to a degree and all show similar behaviour corresponding to excitation of surface and collective nuclear excitations. In this range ^{3}He reactions share with ^{4}He reactions the function as probes for further study of these surface and collective phenomena.

Further results relevant to the determination of the reaction mechanisms involved in ^{3}He studies will be presented in the detailed survey of the experimental data presented in § 4 of this review.

§3. Nuclear Spectroscopy

3.1. *Introduction*

In this section ^{3}He reaction measurements which have been designed to yield information regarding individual nuclear energy levels will be presented as distinct from the experiments described in the previous section (§2) where the reaction mechanism itself was investigated. The advantages of ^{3}He as a projectile in nuclear spectroscopy have already been mentioned in §1. The most important are the very high Q values which permit the formation and study of highly excited states of the residual nuclei and the high charge-to-mass ratio which allows the production of proton-rich nuclides otherwise inaccessible to study by precision techniques. In general very little detailed information about the levels of the compound state formed by ^{3}He capture is obtained because the excitation is sufficiently high that individual levels overlap. Distinct resonance structure is unusual in the excitation functions and the few exceptional cases that have been amenable to resonance analysis have already been discussed in §2. These special nuclei are listed in table 3 of §2.2.1, together with the level parameters derived from the resonance fits to the experimental data. In view of its limited application to ^{3}He studies no further discussion of this technique will be presented.

3.2. *Particle Spectroscopy*

Much of the detailed information on nuclear energy levels which results from ^{3}He investigations pertains to excited states of the residual nuclei formed in reactions of the type (^{3}He, p), (^{3}He, n), (^{3}He,α) and (^{3}He, d). Precision measurements of the charged particle energies using magnetic spectrographs lead to accurate values of the excitation energies of the corresponding levels in the product nuclei. See for example the measurements of Hinds and Middleton presented in figure 30. Similar precision and energy resolution are not attainable in neutron measurements, but even the lower accuracies which are attainable produce results of interest because (^{3}He, n) reactions often lead to proton-rich nuclei which cannot readily be studied via other reactions. For example, the most accurate information on ground state masses and energy levels of ^{14}O and ^{12}N was obtained from study of (^{3}He, n) reactions (Bromley *et al.* 1957, Ajzenberg-Selove, Bullock and Almqvist 1957 a, b).

There is further interest in comparing (^{3}He, p) and (^{3}He, n) reactions from the same target since they lead to different members of isobaric multiplets. If forces between nucleons are independent of their charge, levels of the same properties will occur in different isobars and definite relationships (Adair 1952) exist between the cross sections for (^{3}He, p) and (^{3}He, n) reactions leading to such analogue levels. Although Coulomb effects cause deviations from the predictions based on the assumption of charge independence there is evidence for a definite correspondence between the cross sections to corresponding levels. An example of this is shown in figure 31 where the spectra from the ^{9}Be(^{3}He, p)^{11}B and the ^{9}Be(^{3}He, n)^{11}C (Almqvist *et al.* 1956, 1960, Stewart, van de Zwan and Almqvist 1959, unpublished) reactions are compared. Other examples are provided by the reactions ^{12}C(^{3}He, n)^{14}O and ^{12}C(^{3}He, p_1)^{14}N (2·31 Mev) (Bromley *et al.* 1957) and ^{10}B(^{3}He, n)^{12}N and ^{10}B(^{3}He, p)^{12}C (15·1 Mev) (Ajzenberg-Selove, Bullock and Almqvist 1957 a, b, Almqvist *et al.* 1959) which in each case show very similar angular

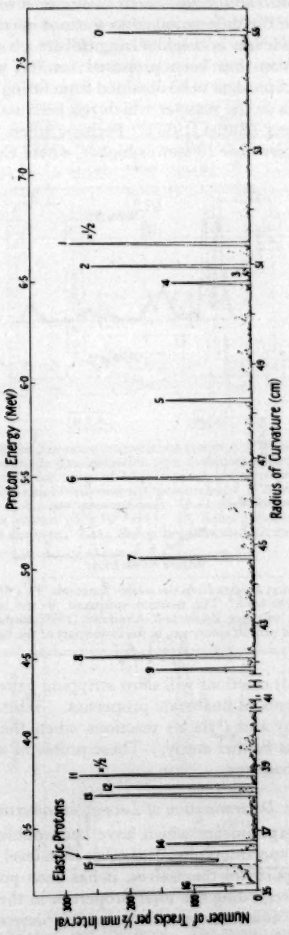


Figure 30. Proton spectrum from the bombardment of ^{16}O with 5·70 MeV ^{3}He ions, using magnetic deflection and photographic plate detection. An absolute accuracy of ± 10 kev is quoted for the peak locations and hence for the excitation energies of levels in the residual nucleus ^{19}F. From Hinds and Middleton (1959 a).

distributions for protons and neutrons to analogue levels (see §§ 4.9 and 4.12). This correspondence in the differential cross sections may be useful in determining which are the analogue levels in neighbouring isobars.

As yet no mechanism has been proposed for ^{3}He reactions which permits unambiguous spectroscopic data to be obtained from fitting the angular distributions of the charged particles in the manner which has been so successful for (d, p) and (d, n) stripping reactions (Butler 1957). Perhaps when more systematic studies are carried out at energies near 10 MeV or higher, where Coulomb effects are small,

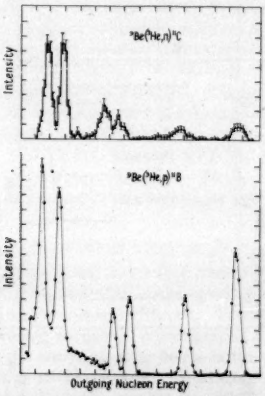


Figure 31. Neutron and proton spectra from the mirror reactions ^{9}Be (^{3}He, n) ^{11}C and ^{9}Be (^{3}He, p) ^{11}B measured at $E_{^3He} = 2{\cdot}0$ MeV. The neutron spectrum, in the upper part of the figure, was obtained by Stewart, van der Zwan and Almqvist (1959, unpublished) using photographic plate detectors, and the proton spectrum, in the lower part of the figure, by Almqvist *et al.* (1956) using a CsI detector.

then (^{3}He, α) and (^{3}He, d) reactions will show stripping patterns which are amenable to simple analysis in terms of final state properties. Whether or not this will also be the case for (^{3}He, p) and (^{3}He, n) reactions when their mechanism becomes better understood awaits further study. These problems are discussed more fully in § 2 on Reaction Mechanisms.

3.3. *Determination of Level Characteristics*

Although in most experiments which have been published to date it has not been possible to extract spectrographic data other than level energies from measurements on the reaction particles themselves, it has been possible to obtain a large amount of information regarding the level properties in the final nuclei by using a gamma ray detector in coincidence with a particle detector in the arrangement shown in figure 32. With this arrangement a particular level in the residual nucleus may be chosen for study by selecting the corresponding particle group with the voltage window of a single channel analyser and examining with the second detector

the radiations that are in time coincidence with pulses in this window. As an example, the shaded area in figure 33 (Almqvist *et al.* 1959) shows a window set on the proton group from the ^{10}B (^{3}He, p) ^{12}C reaction and below this is the coincident pulse spectrum from the gamma ray detector. From such data the de-excitation

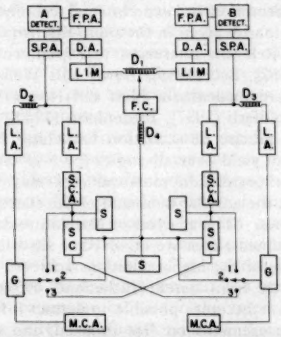


Figure 32. Block diagram of the fast–slow coincidence system. The occurrence of a coincidence is determined by a biased diode discrimination type coincidence circuit (Bell, Graham and Petch 1952). Coincidence resolving times of 10–30 mμ sec are conveniently used. The slow linear signals are used for pulse-height selection and measurement. The abbreviations used are : S.P.A., slow preamplifier ; F.P.A., fast preamplifier ; D.A., distributed amplifier ; LIM, limiter ; F.C., fast coincidence circuit ($2\tau \simeq 10^{-8}$ sec) ; D, delay ; S.C.A., single channel analyser ; M.C.A., multichannel analyser ; L.A., linear amplifier and G, gate.

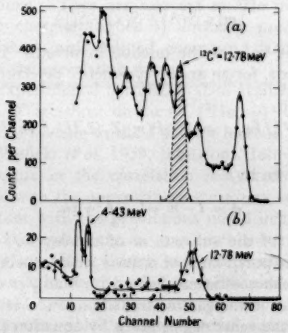


Figure 33. (*a*) Proton groups from ^{10}B (^{3}He, p) ^{12}C* in the region of 12·78 MeV excitation detected with a CsI scintillator at 90° to 2 MeV ^{3}He beam. The shaded area indicates the pulse heights accepted by the voltage window. (*b*) The distribution of gamma-ray pulses which are coincident with the particle pulses accepted by the window. Only the 12·7 MeV ground state transition is seen ; the 4·43 MeV peak was shown to be due to random coincidences. From Almqvist *et al.* (1959).

branching of the selected state may be obtained; in the case illustrated the 12·78 Mev state in ^{12}C was shown to de-excite 2% by gamma ray transitions to the ground state and 98% by alpha particle emission. Other examples of particle–gamma and gamma–gamma coincidence measurements are shown in figure 34 (Bromley *et al.* 1959 e) and figure 17 (Kuehner *et al.* 1959). A more detailed discussion of both the theoretical and experimental techniques of angular correlation experiments using ^3He ions is given in the more extensive review article (Bromley and Almqvist 1959). Relevant references to experimental techniques are Bell, Graham and Petch (1952), Bell (1955), Lewis and Wells (1954); references to theoretical techniques are Biedenharn, Blatt and Rose (1952), Rose (1953), Rose (1958), Devons and Goldfarb (1957), Biedenharn (1960), and Goldfarb (1959).

In order to obtain precise de-excitation branching values it is necessary to integrate the gamma ray yield over all angles for a given position of the particle counter. Such angular correlation measurements may in certain special cases yield information about the angular momenta of the states involved and about the gamma ray multipolarities. The analyses of correlation data in general is simplest if only isolated, well defined states are involved, a situation which rarely pertains to ^3He reactions because of the high excitations achieved in the compound state. However, it has recently been noted (Litherland 1959) that there are special circumstances in which it becomes possible to extract information from particle–gamma correlation measurements on ^3He induced, and other, reactions without requiring information regarding the characteristics of the compound system. Litherland (1959) has noted that in a triple correlation of the A(x, yγ) B form, for example, if the reaction particles y are detected at 0° or 180° to the direction of the incident beam of particles x, the only magnetic substates m of the excited state J of the final nucleus, which contribute to the correlation measured between the particles y, so detected, and the associated gamma radiation are those satisfying

$$|m| \leqslant S(\mathrm{A}) + S(\mathrm{x}) + S(\mathrm{y}) \qquad \ldots\ldots(21)$$

where S is the intrinsic spin of the nuclide involved. The angular correlation, relative to the x beam axis, for an axially symmetric detector, is predicted in general to be:

$$W_{LL'}(\theta) = \sum_m \sum_k (-1)^{\phi} (JJm, -m|k0)\, P(m)\, Z_1(LJL'J, Ik)\, P_k(\cos\theta) \qquad \ldots\ldots(22)$$

where the phase ϕ is given by

$$\phi = I - m + L + L' + \tfrac{1}{2}k. \qquad \ldots\ldots(23)$$

$P(m)$ is the population of the substate m of the level J, hence $P(m) = P(-m)$. $P_k(\cos\theta)$ is the Legendre polynomial of order k with respect to the x beam axis, and Z_1 is the associated Racah coefficient defined by Sharp *et al.* (1954), L and L' are the interfering multipoles in the gamma transition, and I is the final state spin.

In situations where the value of $|m|$ given by equation (21) is zero or $\frac{1}{2}$, e.g. for (α, pγ) or (^3He, αγ) reactions on spin zero targets the angular correlation is uniquely defined by the Z_1 coefficient in terms of angular momenta and multipole mixtures involved in the gamma transition. The fact that particle detectors of finite aperture are required implies that small admixtures of substates differing by ±1

can contribute to the correlations. This contribution may be estimated in detail from the correlation formalism with detector geometries normally in use; such corrections are often quite negligible (Litherland and McCallum 1960).

In cases such as (^{3}He, pγ) or (^{3}He, nγ), the ratio $P(1)/P(0)$ appears as an additional parameter in the predicted correlation in equation (22). If the situation is such that both $P_2(\cos\theta)$ and $P_4(\cos\theta)$ intensities are obtained from the experimental correlation a unique determination of J, of the multipole mixture, as well as of this ratio is possible from a single measurement. If only the $P_2(\cos\theta)$ intensity is obtained, a measurement of the linear polarization of the gamma radiation (Fagg and Hanna 1959) suffices not only to fix all parameters but also determines the parity change in the gamma transition.

Clearly the utility of this technique decreases rapidly with increasing channel spins and consequent increasing number of parameters to be determined.

An extension of these ideas (Litherland and Ferguson 1960) has been made to situations where the intermediate particle radiation is unobserved and measurements are carried out on the correlation between cascade gamma radiations de-exciting states in the final nucleus (see also Warburton and Rose 1958). By measuring the correlations in a number of geometries it is possible to extract spectroscopic information on the level assignments and transition multipolarities as well as the population parameter ratios for the upper state (Litherland and Ferguson 1960). This latter technique requires that the individual gamma radiations be discernible in the direct reaction spectrum; this situation is somewhat anomalous in ^{3}He induced reactions because of the high Q values and large numbers of open channels. This technique has been used in a study of the ^{24}Mg(pp$'\gamma\gamma$)^{24}Mg reaction (Batchelor *et al.* 1960).

These are powerful techniques in that their application is entirely independent of the reaction mechanisms involved and depends only on general angular momentum considerations. They are relevant to ^{3}He induced reactions in that the large ^{3}He binding energies (figure 1) normally preclude any possibility of correlation studies through well defined compound states which are otherwise required for analysis of the correlation data.

The only relevant experimental correlation data available to date with particles detected at other than 90° are those on the ^{13}C(^{3}He, $\alpha\gamma$)^{12}C, ^{11}B(^{3}He, d, γ)^{12}C, and ^{10}B(^{3}He, pγ)^{12}C reactions leading to the 4·43 MeV level in ^{12}C (Holmgren, Wolicki and Johnston 1959 a, Wolicki *et al.* 1959, Johnston, Holmgren and Wolicki 1959). The very strong structure in the correlation measured with the alpha particle counter at 10° together with the approximate symmetry of the correlation relative to $\theta_\gamma = 90°$ are consistent with the predictions noted in the previous paragraphs. The deviations from exact symmetry are presumably attributable to the fact that the θ_p is not exactly 0° or 180°. Much more experimental information of this type is needed to throw light on the reaction mechanism (see §2.3) as well as to obtain spectroscopic data.

Another circumstance in which gamma–gamma correlations are amenable to simple analyses is that where the magnetic substates of the de-exciting level are normally populated. When such an hypothesis is supported by measurements such as those described in §2.2.3, the usefulness of the reaction from the nuclear spectroscopic point of view is greatly enhanced in that a quadrupole correlation

(^{3}He − x − γ − γ) reduces to an effective double γ–γ correlation. In the first case, analysis of correlation data to yield angular momentum and parity assignments to nuclear levels involved and multipole mixture data on the radiations themselves is possible only under very restricted circumstances and with considerable complexity of analysis (Sharp *et al.* 1954, Bromley *et al.* 1957 a); in the latter case the analysis is trivial using any of the excellent tables of functions available for such analysis (e.g. Sharp *et al.* 1954, Ferguson and Rutledge 1957). This technique should be

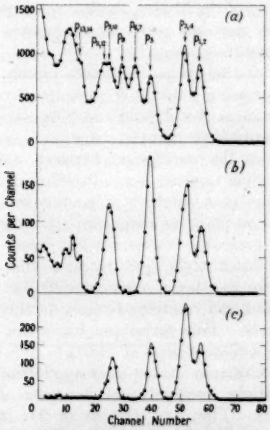


Figure 34. Proton spectra from the ^{14}N (^{3}He, p) ^{16}O reaction for $E_{^3He}$ = 2·1 MeV. (*a*) Direct proton spectrum as measured with a CsI detector; (*b*) proton spectrum measured in coincidence with all gamma radiation such that $E_\gamma \gtrsim 700$ kev; (*c*) a similar coincidence spectrum with $E_\gamma \geqslant 4{\cdot}5$ MeV. The subscripts in (*a*) identify the corresponding states in ^{16}O. From Bromley *et al.* (1959 e).

amenable to wide exploitation in all reactions where adequate compound system averaging is involved. An example is the ^{14}N (^{3}He, pγ) ^{16}O reaction studied by Kuehner *et al.* (1959).

In addition to measurements of relative radiation and particle widths such as are obtained from the de-excitation branching studies described above it is of course of great importance to measure absolute widths or lifetimes. A relevant example is the measurement by Temmer and Heydenburg (1958) who determined the lifetimes of the 666 and 593 kev states in ^{22}Na found in the reaction ^{20}Ne (^{3}He, p) ^{22}Na to be 0·014 and 0·266 microseconds respectively. A survey of techniques of lifetime measurements, with special reference to ^{3}He reactions, is given by Bromley and Almqvist (1959).

Other relevant references to spectroscopic studies using ^{3}He are given in § 4 which surveys the published experimental ^{3}He data. In particular coincidence techniques have been extensively used to study levels of the residual nuclei formed by ^{3}He bombardment of ^{9}Be, ^{10}B, ^{11}B, ^{12}C, ^{14}N, ^{16}O and ^{20}Ne.

3.4. *Coulomb Excitation Measurements*

Only one instance of the use of the ^{3}He induced Coulomb excitation has been reported (Bromley, Kuehner and Almqvist 1959, 1959 a). Coulomb excitation of each of the 110 and 197 Mev levels of ^{19}F, the 446 Mev level of ^{23}Na, the 160 Mev level of ^{47}Ti and the 126 Mev level of ^{55}Mn has been examined in the incident energy range from 1·0 to 2·0 Mev. Thick target excitation curves for both the single E1 case (110 Mev level in ^{19}F) and the E2 cases are in good agreement with the theoretical predictions (Alder *et al.* 1956) as modified for the thick targets used (Lindhard and Scharff 1953). The possibility of augmented cross sections in ^{3}He induced Coulomb excitation resulting from the relatively large expected polarizability of the ^{3}He ion has been examined. Such polarization effectively mixes E_λ and $E_{\lambda+3}$ excitations. It was found that in the case of $\lambda = 1$, for example, the ratio of the cross section contribution of the higher multipole to that of the lower is given approximately by $10^{-3}\alpha$ where α is the ^{3}He polarizability in units of the ^{3}He radius. Since it would be anticipated that $\alpha \ll 1$, it follows that the polarization effect is of negligible importance in normal Coulomb excitation measurements (McManus as quoted in Bromley, Kuehner and Almqvist 1959 a).

Use of ^{3}He together with ^{4}He incident ions has particular advantages in allowing the determination of transition multipolarities from measurements of the relative Coulomb excitation yields at classically corresponding energies, i.e. for the same values of the adiabatic parameter

$$\xi = \frac{Z_1 Z_2 e^2}{\hbar}\left(\sqrt{\frac{M}{2}}\frac{\sqrt{E_i}-\sqrt{E_f}}{\sqrt{(E_i E_f)}}\right) \qquad \ldots\ldots(24)$$

where M is the reduced mass of the incident ion, $Z_1 e$ its charge, E_i and E_f its energy before and after interaction respectively (Bjerregaard and Huus 1954). This technique avoids the necessity of determining target thickness, counter efficiencies or absolute beam flux. For equal ξ values, however, the ranges of the incident particles are not equal and the yields must be corrected for the greater effective target thickness seen by the longer range particle in a thick target. Since the relationship between the ranges of protons and alpha particles is not well established (Whaling 1958), ratio measurements using these ions have been restricted to thin targets. Because the ranges of ^{3}He and ^{4}He ions are uniquely related (Whaling 1958)

$$R_{^3\mathrm{He}}(AE) = aR_{^4\mathrm{He}}(E/a) \qquad \ldots\ldots(25)$$

where

$$a = M_{^3\mathrm{He}}/M_{^4\mathrm{He}}$$

it is possible, using these ions, to use thick targets with much greater yield.

Figure 35 shows the results of measurements on a Nb F_5 target at the indicated energies (Bromley, Kuehner and Almqvist 1959, 1959 a). The much greater bremsstrahlung yield associated with ^{3}He bombardment is clearly evident in this figure. It should be noted that the ^{3}He induced bremsstrahlung yield does not necessarily increase with increasing target number. In the case illustrated, for example, the yield from fluorine ($Z = 9$) is several times that from niobium ($Z = 41$) simply because the radial wave function overlap is much more favourable in the former case, i.e. since the ^{3}He ion approaches much more closely to the fluorine target nucleus. The matrix element factor which depends upon this overlap (Alder *et al.* 1956) then more than compensates for the Z dependence.

Table 8 lists the ratio results obtained. As indicated, the agreement between theory and experiment is quite acceptable. In the case of ^{23}Na the large quoted error reflects the fact that the transition energy is close to the ubiquitous 0·511

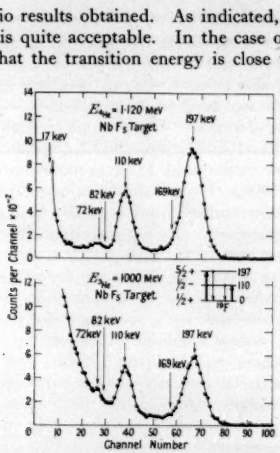


Figure 35. Gamma radiation spectra from Coulomb excitation of the low levels of ^{19}F using ^{3}He and ^{4}He ions of classically corresponding energy. The inset level diagram shows the transitions involved. 169 and 82 kev are the residual energies following escape from the 1½ in. × 1 in. NaI crystal of the 33 kev iodine K x-ray; 72 and 17 kev are the K x-ray energies of lead and niobium respectively.

In extracting the yield of 110 kev radiation from these data an experimentally determined correction equal to 3·3 ± 0·6% of the 197 kev total absorption peak was subtracted to take into account the fact that the Compton quanta from an incident 197 kev radiation scattered at 180° also has an energy of 110 kev. From Bromley, Kuehner and Almqvist (1959, 1959 a).

annihilation radiation associated with the positron decay of proton-rich nuclides produced in ^{3}He reactions. In the case of the E1 transition in ^{19}F, which is the only E1 transition studied by this technique to date, the apparent decrease in the yield ratio with increasing energy is attributable to the onset of nuclear inelastic

Table 8. Coulomb Excitation Yield Ratio Measurements

Target	E_{ex} (kev)	$(J\pi)_i$	$(J\pi)_e$	$E_{^3He}$	$E_{^4He}$	ξ	Yield ratio exp.	Yield ratio theor.
^{19}F	110	1/2+	1/2−	1·000	1·127	0·347	0·76 ± 0·07	0·76
				1·200	1·352	0·230	0·71 ± 0·04	
				1·400	1·578	0·203	0·72 ± 0·13	
				1·750	1·975	0·143	0·59 ± 0·11	
	197	1/2+	5/2+	1·000	1·120	0·681	0·62 ± 0·03	0·64
				1·200	1·347	0·500	0·61 ± 0·02	
				1·400	1·574	0·387	0·63 ± 0·02	
				1·750	1·970	0·269	0·61 ± 0·02	
^{23}Na	439	3/2+	5/2+	1·400	1·550	0·681	0·65 ± 0·03	0·63
				2·000	2·218	0·387	0·63 ± 0·03	
^{55}Mn	126	5/2−	7/2−	1·400	1·549	0·591	0·63 ± 0·03	0·63
				2·000	2·216	0·338	0·62 ± 0·03	

scattering of alpha particles via the 110 MeV level. It would be anticipated that since the number of reaction channels open at any given energy is much greater for ^{3}He than for ^{4}He incident ions, any particular channel, in this case inelastic scattering, would show considerably greater yield for ^{4}He bombardment at low energies. Such yield would result in decrease in the yield ratio. The appearance of such effects in the transition from the 110 MeV state and not in that from the 197 MeV state is consistent with the higher excitation and higher angular momentum of the higher state.

§4. Survey of Experimental Results

Excellent compilations of the results of nuclear reaction studies for $A \lesssim 40$ have been published by Ajzenberg-Selove and Lauritsen (1959) and by Endt and Braams (1957); further comprehensive coverage is provided by the National Bureau of Standards (U.S.A.) Nuclear Data Service (Way 1958). It would indeed be pointless to duplicate this presentation herein. It may be argued however that a programme of study of ^{3}He (or ^{3}H) induced reactions is one of the most fruitful remaining to experimentalists working with relatively low energy accelerators; for the convenience of those embarking upon such a programme, with ^{3}He ions, it is considered worth while to present a survey of the experimental results so far obtained.

4.1. *Average Reaction Cross Sections*

Most of the reported work on ^{3}He reactions to date has been carried out with incident energies in the energy range $0{\cdot}8 \leqslant E_{^3\mathrm{He}} \leqslant 10$ MeV. As shown in table 9, the reaction cross sections are typically about 10 mbn resulting in adequate yield for detailed measurements involving coincidence and correlation studies. The

Table 9

Reactions	$E_{^3\mathrm{He}}$ (MeV)	$\langle\sigma\rangle$ (mbn)	Reactions	$E_{^3\mathrm{He}}$ (MeV)	$\langle\sigma\rangle$ (mbn)
D (^{3}He, p)	0·64 (resonant)	700	^{10}B (^{3}He, n) ^{12}N	2·5	5
			^{10}B (^{3}He, p) ^{12}C	2·0	20
D (^{3}He, γ)	0·64 (resonant)	0·05	^{10}B (^{3}He, d) ^{11}C	2·0	10
T (^{3}He, d) ^{4}He	0·8	180	^{11}B (^{3}He, p) ^{13}C	3·0 – 5·0	2
T (^{3}He, pn) ^{4}He	0·2	10	^{11}B (^{3}He, d) ^{12}C	3·0 – 5·0	5
^{3}He (^{3}He, 2p) ^{4}He	0·8	2	^{12}C (^{3}He, n) ^{14}O	2·0 – 3·0	30
	0·2	$2{\cdot}5 \times 10^{-3}$	^{12}C (^{3}He, p) ^{14}N	2·0 – 3·0	20
			^{12}C (^{3}He, α) ^{11}C	2·0 – 3·0	50
^{4}He (^{3}He, γ)	1·0	5×10^{-4}			
	0·5	4×10^{-5}	^{13}C (^{3}He, p) ^{15}N	2·0	2
				4·5	14
^{6}Li (^{3}He, n) ^{8}B	5·0	20	^{13}C (^{3}He, d) ^{13}C	2·0	1
^{6}Li (^{3}He, p) ^{8}Be	2·0 – 5·0	35		4·5	2
			^{14}N (^{3}He, p) ^{16}O	2·0	0·3
^{9}Be (^{3}He, n)	2·0	10			
^{9}Be (^{3}He, p)	2·5	10	^{16}O (^{3}He, p) ^{18}F	2·5 (resonant)	5
			^{16}O (^{3}He, α) ^{15}O	2·5 (resonant)	26

cross section entries which refer in each case to the cross section for a single final state reaction have been obtained in many cases from rough integration of the published differential cross sections. More precise evaluation has not been considered worth while since many of the cross sections have not been measured with great precision and those for reactions leading to different final states in the residual nuclei often vary by a factor of ten or more. These data are only intended as a guide in estimating the magnitude of the yields from ^{3}He reactions; more precise information is available in the following sections where references to the original work are given.

From the point of view of reader convenience, these data will be grouped in somewhat unorthodox fashion according to the target rather than the final nucleus involved. No attempt will be made to provide an exhaustive catalogue of the data; rather it is intended to indicate the scope of the experimental work and to provide detailed references to it and to its interpretation. In each case, following the target heading the compound system and the binding energy of ^{3}He in that system is listed for convenience.

4.2. ^{2}H *Target*

Compound system, ^{5}Li; binding energy, 16·555 MeV

^{2}H(^{3}He, p)^{4}He $Q_m = 18{\cdot}351$ MeV

Extensive measurements on the excitation curve for this reaction have been carried out in many laboratories using both incident ^{3}He ions and deuterons as indicated in table 10. Many of these data are also discussed in §2.2.1.

^{2}H(^{3}He,^{3}He)^{2}H

The elastic scattering of deuterons on ^{3}He has been investigated in the energy ranges $1{\cdot}0 \leqslant E_d \leqslant 3{\cdot}25$ MeV (Brown *et al.* 1953) and $0{\cdot}3 \leqslant E_d \leqslant 0{\cdot}6$ MeV (Freier and Holmgren 1954). Destructive interference between potential and resonant elastic scattering is observed in the vicinity of the 435 keV resonance. See also §2.2.1.

^{2}H(^{3}He, γ)^{5}Li $Q_m = 16{\cdot}555$ MeV

Capture gamma radiation resulting from the de-excitation of the 16·81 MeV level in ^{5}Li has been detected for $0{\cdot}2 \leqslant E_d \leqslant 2{\cdot}85$ MeV. A maximum cross section $\sigma = 50 \pm 10$ μbn was measured for $E_d = 0{\cdot}45 \pm 0{\cdot}04$ MeV. Non-resonant capture was observed at energies above this maximum (Blair, Hintz and Van Patter 1954). The capture gamma radiation is found not to be monoergic but extends over several MeV with an upper energy cut-off of 16·5 MeV (Kunz, Butler and Holmgren 1955 a, b).

4.3. ^{3}H *Target*

Compound system, ^{6}Li; binding energy, 15·790 MeV

^{3}H(^{3}He, d)^{4}He, ^{3}H(^{3}He, p)^{5}He and ^{3}H(^{3}He, np)^{4}He reactions have been studied by Moak (1953) for $0{\cdot}1 \leqslant E_{^3\mathrm{He}} \leqslant 0{\cdot}8$ MeV. These reactions are non-resonant in this energy range. The total reaction cross section for ^{3}He on ^{3}H rises smoothly from 450 μbn at 100 keV to 0·16 bn at 800 keV. The cross sections

for the three reactions are in the ratios 43 : 6 : 51 respectively ; these ratios are constant within 2% over the range studied. Almqvist *et al.* (1951) found ratios of 43 : 3 : 54 at 240 kev.

The calculated Q value for the ^3H (^3He, p) ^5He reaction is 11·136 Mev ; measurements of 11·18 ± 0·07 Mev and 11·13 ± 0·07 Mev have been reported by Almqvist *et al.* (1951, 1953) and by Moak (1953) respectively.

4.4. ^3He *Target*

Compound system, ^6Be ; binding energy, 10·4 Mev

$^3He(^3He, p)^5Li$ $\qquad Q_m = 11{\cdot}062$ Mev

Excitation curves for this reaction have been measured for $100 \leqslant E_{^3He} \leqslant 800$ kev (Good, Kunz and Moak 1951, 1954). In this range the cross section increases monotonically with incident energy. At $E_{^3He} = 200$ kev, $\sigma = 2{\cdot}5$ μbn ; at lower energies the cross section is adequately fitted by a Gamow expression and at higher energies the data require significant contributions from D-wave and higher angular momentum components in the incident beam.

Proton spectra measured at $E_{^3He} = 360$ kev (Good, Kunz and Moak 1951) and 240 kev (Almqvist *et al.* 1953) show a peak corresponding to the ground state transition superposed on a broad continuum from 3-body disintegrations in the ^3He(^3He, 2p)^4He reaction ($Q = 12{\cdot}858$ Mev). No evidence has been found for proton groups of lower energy corresponding to excited states in ^5Li (Good, Kunz and Moak 1954, Almqvist *et al.* 1953).

4.5. ^4He *Target*

Compound system, ^7Be ; binding energy, 1·584 Mev

$^4He(^3He, ^3He)^4He$

A study of the elastic scattering of ^3He on ^4He has been carried out by Miller and Phillips (1958) for $3 \leqslant E_{^3He} \leqslant 5{\cdot}5$ Mev corresponding to excitation in the compound ^7Be in the range $3{\cdot}28 \leqslant E_x(^7Be) \leqslant 4{\cdot}73$ Mev. A resonance corresponding to the second excited state of ^7Be was located at $E_{^3He} = 5{\cdot}17$ Mev with an observed width of 0·180 Mev. These correspond to $E_x(^7Be) = 4{\cdot}53 \pm 0{\cdot}02$ Mev and $\Gamma_{cm} = 0{\cdot}102$ Mev. With an assignment $F_{7/2}$ to this state, the ^3He reduced width is $3{\cdot}0 \times 10^{-13}$ Mev cm^{-1} which is 0·36 of the Wigner limit. The formal resonance energy $E_x = 3{\cdot}43$ Mev (Vogt 1958) and the interaction radius used was $4{\cdot}4 \times 10^{-13}$ cm. The P-wave phase shifts suggest that the ^7Be ground state width is essentially equal to the Wigner limit and the S-wave phase shift indicates hard-sphere scattering with $R = 2{\cdot}8 \times 10^{-13}$ cm (Chiba *et al.* 1958).

$^3He(^4He, \gamma)^7Be$ $\qquad Q_m = 1{\cdot}584$ Mev

The cross section for radiative capture of ^4He by ^3He has been measured by Holmgren and Johnston (1958, 1959) at $E_{^3He} = 0{\cdot}48$, 0·72, 0·94, 1·13 and 1·32 Mev. Figure 36 presents the results of these measurements ; the cross section is adequately represented in this energy range by a Gamow expression of the form

$$\sigma = \frac{2{\cdot}8(1 - 0{\cdot}55\,E_\alpha)}{E_\alpha} \exp(-7{\cdot}92\,E_\alpha^{-1/2})\ \mu\text{bn}. \qquad \ldots\ldots(26)$$

Table 10. $^2H(^3He, p)^4He$

Reference	Projectile energy range	Remarks
1. G. A. Baker, J. T. Holloway, L. D. P. King and R. E. Schreiber, *USAEC Report* AECD–2189 (1948)		
2. J. Hatton and G. Preston, *Nature, Lond.*, **164**, 143 (1949)	$0{\cdot}08 \leqslant E_d \leqslant 0{\cdot}180$	
3. B. H. Flowers, *Proc. Roy. Soc.* A, **204**, 503 (1950)		Theoretical treatment of excitation curve data in reaction–length formalism
4. S. T. Butler and J. L. Symonds, *Phys. Rev.*, **83**, 858 (1951)		Analysis of ^3He (dp) ^4He and T (dn) ^4He angular distributions at $E_d = 10$ Mev in stripping formalism
5. J. C. Allred, *Phys. Rev.*, **84**, 695 (1951)	$E_d = 10{\cdot}2$ Mev	
6. J. L. Tuck, W. R. Arnold, J. A. Phillips, G. A. Sawyer and E. J. Stovall, *Phys. Rev.*, **88**, 159 (1952)	$0{\cdot}035 \leqslant E_d \leqslant 0{\cdot}095$	
7. A. B. Lillie, T. W. Bonner and J. R. Conner, *Phys. Rev.*, **88**, 473 (1952)		
8. R. J. S. Brown, G. Freier, H. D. Holmgren, W. R. Stratton and J. L. Yarnell, *Phys. Rev.*, **88**, 253 (1952)		
9. J. L. Yarnell, R. H. Lovberg and W. R. Stratton, *Phys. Rev.*, **90**, 292 (1953)	$0{\cdot}250 \leqslant E_d \leqslant 3{\cdot}50$	
10. R. G. Jarvis and D. Roaf, *Proc. Phys. Soc.* A, **66**, 310 (1953)	$0{\cdot}019 \leqslant E_d \leqslant 0{\cdot}045$	
11. R. G. Jarvis and D. Roaf, *Proc. Roy. Soc.* A, **218**, 432 (1953)		Gamow barrier penetration factor fits energy dependence of cross section but not its absolute value
12. W. R. Arnold, J. A. Phillips, G. A. Sawyer, E. J. Stovall and J. L. Tuck, *Phys. Rev.*, **93**, 483 (1954)	$E_d \leqslant 0{\cdot}1$	$\sigma = 16$ mbn ($\pm 3\%$) at $E_d = 93$ kev ; $\sigma = 0{\cdot}124$ mbn (-5%) at $E_d = 36$ kev; Gamow shape for $E_d < 100$ kev, i.e. $\sigma = (18{\cdot}2 \times 10^9/E) \times \exp(-91E^{-\frac{1}{2}})$

Reference	Projectile energy range	Remarks
13. G. Freier and H. Holmgren, *Phys. Rev.*, **93**, 825 (1954)	$0{\cdot}3 \leqslant E_d \leqslant 0{\cdot}6$	$\sigma_T = 0{\cdot}94$ barns at resonance maximum $E_d = 450 \pm 40$ kev. Observed interference between potential and resonant scattering amplitudes
14. W. E. Kunz, *Phys. Rev.*, **91**, 462 (1953), **97**, 456 (1955)	$0{\cdot}1 \leqslant E_{^3He} \leqslant 0{\cdot}8$	$\sigma_T = 0{\cdot}695 \pm 0{\cdot}014$ barns at resonance maximum $E_{^3He} = 640$ kev. Angular distribution isotopic at $E_{^3He} = 200, 290, 350$ kev. Compound state $3/2^+$ at $16{\cdot}2 \pm 0{\cdot}3$ Mev in ^{5}Li
15. D. P. Hale and R. D. Present, *Phys. Rev.*, **99**, 1366 (1955)		Analysed excitation curve data, centre-of-mass energies 64–954 kev, using log. derivative resonance formalism
16. A. P. Kliucharev, B. N. Esel'son and A. K. Val'ter, *Doklady Akad. Nauk. S.S.S.R.*, **109**, 475 (1956)		$(d\sigma/d\Omega)_{90°} = 63{\cdot}4$ mbn sterad^{-1}
17. D. L. Booth, R. S. Hill, F. V. Price and D. Roaf, *Proc. Phys. Soc.* A, **70**, 863 (1957)	$0{\cdot}1 \leqslant E_d \leqslant 0{\cdot}5$	$(E_d)_{res} = 435 \pm 30$ kev E_d (kev) — σ_T (barns) 503·2 — $0{\cdot}476 \pm 0{\cdot}008$ 402·6 — $0{\cdot}606 \pm 0{\cdot}050$ 301·9 — $0{\cdot}482 \pm 0{\cdot}040$ 201·3 — $0{\cdot}15 \pm 0{\cdot}01$ 110·7 — $0{\cdot}023 \pm 0{\cdot}001$ Angular distribution isotropic for $E_d \leqslant 500$ kev
18. J. L. Johnson and W. S. Porter, *Phys. Rev.*, **99**, 630A (1955)		Total σ for ^{3}He (dp) ^{4}He fitted with single level Breit–Wigner formalism assuming only S-wave deuterons
19. V. Kistiakowsky-Fischer, *Bull. Amer. Phys. Soc.*, Ser. II, **2**, 28 (1957)	$E_d = 21$	$(d\sigma/d\Omega)_{15°\,(lab)} = 0{\cdot}72 \pm 0{\cdot}28$ mbn sterad^{-1}. Angular distribution of protons not fitted assuming normal interaction radii.
20. W. E. Kunz, J. W. Butler and H. D. Holmgren, *Bull. Amer. Phys. Soc.*, Ser. I, **30**, 25 (1955)[a]		Radiative capture gamma radiation not monoergic but extends over several Mev with 16·5 Mev upper cut-off. Ratio of gamma to proton yield increases with incident energy above the proton resonance.

This is to be compared with the cross section for the mirror reaction $^3H(^4He,\gamma)^7Li$ measured by the same authors as

$$\sigma = \frac{0{\cdot}12(1-0{\cdot}51E_\alpha)}{E_\alpha}\exp(-3{\cdot}96E_\alpha^{-1/2})\,\mu\text{bn}. \qquad \ldots\ldots(27)$$

At $E_{^3He} = 1{\cdot}32$ MeV 50% of the transitions in the $^3He(^4He,\gamma)^7Be$ reaction involve the 0·43 MeV state of ^7Be.

$^3He(^4He,p)^6Li$ $\qquad Q_m = 4{\cdot}023$ MeV

Using 32·5 and 35·5 MeV alpha particles on a ^3He target Chiba *et al.* (1958) have reported observation of proton groups from this reaction.

Additional measurements on this reaction have been reported by Henley (1958) and by Ball, Fairhall and Halpern (1958).

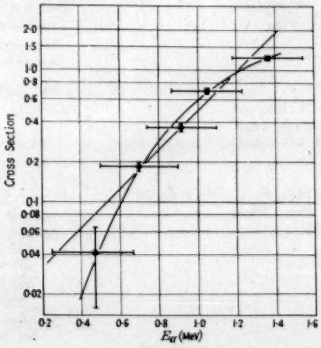


Figure 36. Excitation curve for the $^3He(\alpha\gamma)^7Be$ non-resonant capture reaction. From Holmgren and Johnston (1959).

4.6. ^6Li *Target*

Compound system, ^9B; binding energy, 16·594 MeV

$^6Li(^3He,n)^8B$ $\qquad Q_m = -1{\cdot}616$ MeV

The neutron excitation function for this reaction has been measured for $E_{^3He}$ between 2·9 and 6·0 MeV (Dunning, Butler and Bondelid 1958). The reaction threshold energy was determined absolutely to be $2{\cdot}9661 \pm 0{\cdot}0017$ MeV using an electrostatic analyser (see figure 37). Two additional thresholds were located at $3{\cdot}9 \pm 0{\cdot}1$ and $4{\cdot}16 \pm 0{\cdot}05$ MeV bombarding energy corresponding to excited states in ^8B at excitations of $0{\cdot}6 \pm 0{\cdot}1$ and $0{\cdot}80 \pm 0{\cdot}05$ MeV with widths $\Gamma = 0{\cdot}2 \pm 0{\cdot}1$ and $0{\cdot}07 \pm 0{\cdot}04$ MeV respectively. The neutron yield increases monotonically with energy over the range studied. Farmer and Class (1959) have measured the excitation function for the ground state reaction by observing the positron decay;

the excitation function is linearly dependent upon energy for $3{\cdot}4 \leqslant E_{^3\mathrm{He}} \leqslant 5{\cdot}5$ MeV and at 5·0 MeV the cross section is measured to be 20 mbn. These authors, using the counter ratio technique, quote an excitation energy of $0{\cdot}789 \pm 0{\cdot}030$ kev for the excitation energy of the first excited state in ^{8}B. The half-life for the positron decay of the ground state was determined to be $0{\cdot}84 \pm 0{\cdot}02$ second.

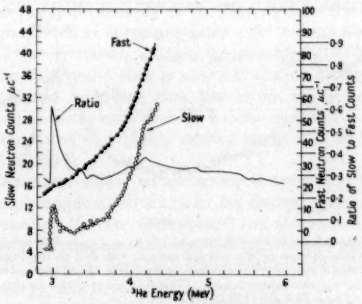


Figure 37. Neutron excitation and counter ratio curves measured with the ^{6}Li (^{3}He, n) ^{8}Be reaction. From Dunning, Butler and Bondelid (1958).

^{6}Li (^{3}He, p) ^{8}Be $\qquad Q_m = 16{\cdot}779$ MeV

Differential excitation functions for the reactions leading to the ground and 2·49 MeV states of ^{8}Be have been measured for incident ^{3}He energies with range from 0·8 to 5·1 MeV (Moak and Wisseman 1956, Schiffer *et al.* 1956 a, b).

Table 11

E_{res} (MeV)	Γ_{res} (MeV)	E_{ex} (MeV)
1·6	0·25	17·6
3·0	1·5	18·6

Resonance structure was observed in these functions superposed on a monotonically increasing background; table 11 lists the observed resonance energies and the corresponding excitation energies in the ^{9}B compound system. It has been suggested (Dunning, Butler and Bondelid 1958) that the observed decrease in the excitation curve above 3·0 MeV may reflect the opening of the neutron channel at 2·966 MeV rather than the presence of a compound system level at this energy. No evidence has been found for excited states in ^{8}Be between 2·94 and 14 MeV excitation with intensities greater than 2% of the first excited state. Angular distributions have been measured at $E_{^3\mathrm{He}} = 1{\cdot}0$, 1·4, 2·0, 3·0, 4·0 and 5·0 MeV (Schiffer *et al.* 1956 a, b). The ground state distributions show pronounced structure and backward maxima suggestive of heavy particle stripping—see § 2.3.1.1;

those to the 2·94 MeV state are approximately symmetric relative to 90° suggesting a compound nucleus interpretation.

Measurements on the proton spectrum from this reaction have been carried out at 800 kev by Kunz, Moak and Good (1953) who advance preliminary evidence for a level in ^{8}Be at an excitation of 12·3–12·6 MeV. Figures 38 (Bigham 1954)

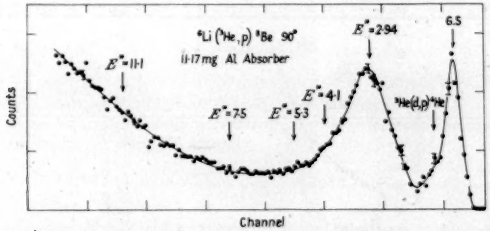


Figure 38. Proton spectrum from the bombardment of ^{6}Li at $E_{^3He}$ = 900 kev obtained using a NaI(Tl) spectrometer. Only groups to the ground state of ^{8}Be and to the broad state at 2·94 MeV are observed. The slight shoulder on the low energy side of the ground state group results from the ^{3}He (d, p) ^{4}He contaminant reaction. There is no evidence in this spectrum for a state at 5·3 MeV in ^{8}Be. From Bigham (1954).

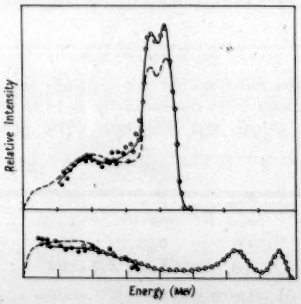


Figure 39. The upper curve is the alpha-particle spectrum from bombardment of ^{6}Li by ^{3}He ions. The open circles were measured with a magnetic spectrometer ; the closed circles with an argon-filled proportional counter. The double humped structure was observed at several angles and energies in the range $600 \lesssim E_{^3He} \leqslant 900$ kev. Assuming the reaction is ^{6}Li (^{3}He, α) ^{5}Li, the measurements show that the ^{5}Li ground state is unstable against proton emission by 1·96 MeV. The second peak corresponds to a level at 1·3 MeV excitation in ^{5}Li. The lower curve shows proton spectrum from ^{5}Li bombarded by ^{3}He. Both spectra at 90° to 900 kev ^{3}He beam. From Almqvist (1954).

and 39 (Almqvist 1954) show proton spectra from ^{3}He on ^{6}Li at 900 kev. The rise in the proton distribution at low energies is probably caused by break-up protons from ^{5}Li formed in the reaction ^{6}Li (^{3}He, α) ^{5}Li. No evidence is found for excited states of ^{8}Be above 2·94 MeV (Almqvist, Allen and Bigham 1955).

^{6}Li(^{3}He, α)^{5}Li $Q_m = 15{\cdot}104$ MeV

This reaction has been studied by Almqvist (1954) and Almqvist, Allen and Bigham (1955) at $E_{^3He} = 900$ kev. Prominent alpha groups of energy 8·3 and 7·5 MeV are observed superimposed on a lower energy continuum of alpha particles which decreases in intensity for $E_\alpha < 3$ MeV. Figure 39 shows a typical alpha particle spectrum from this reaction at a bombarding energy of 900 kev. On the assumption that the reaction proceeds as ^{6}Li(^{3}He, α)^{5}Li$\rightarrow^4$He + p, the prominent alpha groups are interpreted as corresponding to the ground state of ^{5}Li and to an excited state at $\sim$1·3 MeV excitation. From the measured Q values Almqvist (1954) has obtained that the ^{5}Li ground state is unstable against break-up into ^{4}He + p by $1{\cdot}96^{+0{\cdot}10}_{-0{\cdot}20}$ MeV. It should be noted that the lower energy alpha group, while tentatively correlated with the existence of a state in ^{5}Li at $\sim$1·3 MeV excitation, could result from the decay of a highly excited sharp state in ^{8}Be although this is improbable (Almqvist 1954).

Assuming that the large alpha particle group is from ^{6}Li(^{3}He, α)^{5}Li$\rightarrow^4$He + p and assuming isotropic angular distributions the energy spectrum of the break-up protons from the recoiling ^{5}Li can be computed and shown to have an end point of 8·4 MeV. The proton continuum between 8·4 and 11 MeV may be assumed due to direct three-body break-up into two alpha particles and a proton to yield an elliptical proton energy distribution extending from 0 to 16 MeV. The broken curve (figure 39) which is the result of adding the computed spectrum of ^{5}Li break-up protons to the elliptical continuum is in good agreement with the measurements. The continuum of alpha particles below 7 MeV is also consistent with alpha particles from break-up of recoiling ^{5}Li and three-body reaction together with alpha particles from the break-up of the broad 2·9 MeV state in ^{8}Be. The break-up alpha particles from the ground state of ^{8}Be are below 1·4 MeV in energy and are not observed. This analysis assumes relative intensities of 13% via (^{3}He, p)^{8}Be*, 63% via (^{3}He, α)^{5}Li and 24% via (^{3}He, p)2^{4}He (Almqvist 1954). Whether the last reaction is direct three-body break-up or via very broad states in ^{8}Be makes no perceptible difference to the analysis.

Experimentally it was found that at 90° and $E_{^3He} = 700$ kev the ratio of all protons to all alpha particles was $2{\cdot}4 \pm 0{\cdot}2$ which is consistent with the fact that the alpha particles from break-up of the ground state of ^{8}Be were too low in energy to be observed.

4.7. ^{7}Li *Target*

Compound system, ^{10}B; binding energy, 17·780 MeV

^{7}Li(^{3}He, n)^{9}B $Q_m = -9{\cdot}342$ MeV

Neutron spectra from this reaction at a bombarding energy of 2·34 MeV have been measured at 10 angles (Lemonick, Cornwell and Almqvist 1959); 5 excited states have been located at excitations less than 9 MeV in ^{9}B.

^{7}Li(^{3}He, p)^{9}Be $Q_m = 11{\cdot}194$ MeV

Measurements on the proton spectra resulting from 1·25 MeV ^{3}He bombardment of ^{7}Li at 22·5° intervals from 45° to 135° of laboratory angle (Moak *et al.* 1958) are consistent with the existence of levels in ^{9}Be at $1{\cdot}83 \pm 0{\cdot}04$, $3{\cdot}10 \pm 0{\cdot}04$,

$4{\cdot}74 \pm 0{\cdot}08$ MEV and $9{\cdot}1 \pm 0{\cdot}2$ MEV with widths of 0·4, 0·3, $1{\cdot}25 \pm 0{\cdot}25$ and 1·2 MEV respectively in addition to that at $2{\cdot}428 \pm 0{\cdot}005$ which was used for calibration. The experimental data on the 1·83 MEV level are also consistent with a three-body break-up reaction mechanism leading to $^8Be + n + p$: For a fuller discussion of this point see Marion (1960).

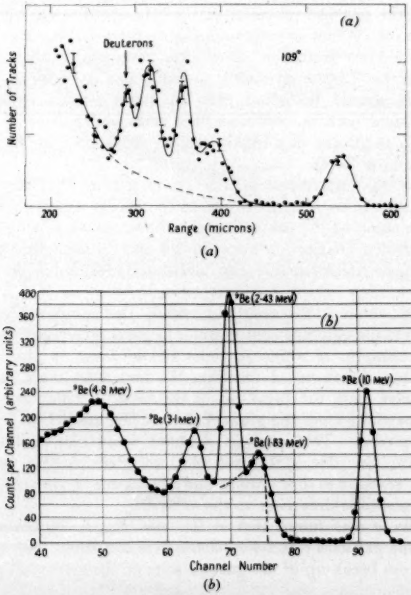


Figure 40. Long range particles from ^7Li bombarded by ^3He. (*a*) Range distribution in a photographic emulsion. The deuteron group is from ^7Li (^3He, d) ^8Be ground state and the proton groups are from ^7Li (^3He, p) ^9Be. From Almqvist, Allen and Bigham (1955). (*b*) Energy distribution of protons from the ^7Li (^3He, p) ^9Be reaction measured with a NaI scintillation spectrometer. From Moak *et al.* (1958).

Charged particle spectra from this reaction have been examined at 720 kev (Moak, Good and Kunz 1954) and 900 kev (Almqvist, Allen and Bigham 1955). Levels are reported at $1{\cdot}73 \pm 0{\cdot}10$, $2{\cdot}40 \pm 0{\cdot}08$, $3{\cdot}05 \pm 0{\cdot}08$, $4{\cdot}74 \pm 0{\cdot}10$ MEV ; the widths of the last two states are 250 ± 100 kev and $1{\cdot}1 \pm 0{\cdot}2$ MEV respectively (Almqvist 1954). Figure 40 (*a*) and (*b*) illustrates the range distribution in photographic emulsions (Bigham 1954) and energy distribution (Moak *et al.* 1958) ; figure 41 shows low energy protons and deuterons measured with a magnetic spectrometer (Almqvist 1954).

^{7}Li(^{3}He, α)^{6}Li $\qquad Q_m = 13{\cdot}320$ MeV

Measurements on this reaction with $E_{^3He} = 900$ kev (Allen, Almqvist and Bigham 1955) locate the levels in ^{6}Li shown in table 12 together with the corresponding experimentally observed widths where available and the deduced level assignments from this and other work (Ajzenberg-Selove and Lauritsen 1959).

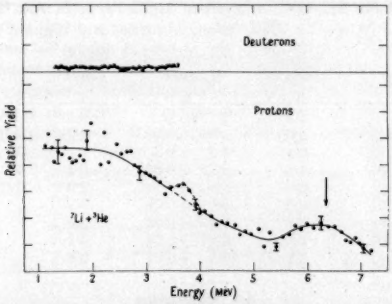


Figure 41. Energy distribution of protons below 7 MeV and deuterons below 3·5 MeV from ^{7}Li bombarded by 900 kev ^{3}He ions. The broad proton group at 6·3 MeV corresponds to a level in ^{9}Be at 4·74 MeV excitation and 1·1 MeV wide. The deuteron continuum may be either associated with broad states in ^{8}Be or with direct break-up into deuterons and two alpha particles. From Almqvist (1954).

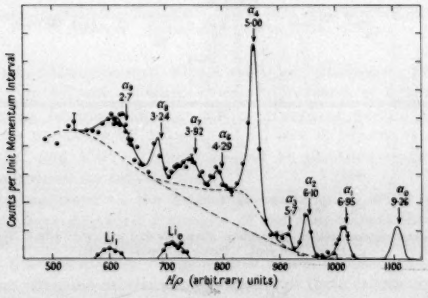


Figure 42. Alpha particle spectrum from the ^{7}Li (^{3}He, α) ^{6}Li reaction measured at 90° to an $E_{^3He} = 900$ kev beam. The chain line is an estimate of the alpha continuum distribution from secondary break-up reactions. The difference between this curve and the broken curve represents the alpha-particle group to the broad ($\Gamma = 1{\cdot}2$ MeV) state at $4{\cdot}8 \pm 0{\cdot}2$ MeV in ^{6}Li. The ^{6}Li recoils correspond to the ground state and 3·56 MeV excited state. The data were taken with a magnetic spectrometer having a line width of 0·75% in momentum. From Bigham (1954).

Figure 42 (Almqvist 1954, Allen, Almqvist and Bigham 1955) shows the alpha particle and ^{6}Li recoil spectrum from ^{7}Li(^{3}He, α)^{6}Li. The observation of ^{6}Li

recoil ions with recoil energy corresponding to their formation in the 3·56 MeV state identifies it as a gamma-emitting level.

At $E_{^3He} = 2{\cdot}1$ MeV the angular distributions of all the resolved alpha-particle groups were found to be isotropic in the centre-of-mass system at this bombarding energy (Bromley *et al.* 1958, unpublished).

Table 12. Results of Magnetic Analysis of Alpha Particles from ^{3}He Bombardment of ^{7}Li (Almqvist 1954, Allen, Almqvist and Bigham 1955)

E_α (MeV)	Relative intensity	Excitation in ^{9}Li (MeV)	Width (MeV)	J	T
8·26	—	0	—	1+	0
6·95	1·0	2·178	<0·1	3+	0
6·10 ± 0·03	1·0	3·56 ± 0·06	<0·1	0+	1
5·7 ± 0·12	0·4	4·3 ± 0·2		2+	0
5·00 ± 0·06	4·1	5·35 ± 0·07	<0·1	2+	1
4·8 ± 0·2	10	5·6 ± 0·2	2·0	1+	0
4·28 ± 0·05	1·0	6·63 ± 0·08	0·2		(1)
3·82 ± 0·05	3·1	7·40 ± 0·08	0·6		(0)
3·24 ± 0·05	1·8	8·37 ± 0·08	0·2		(1)
(2·7 ± 0·1)	3·5	(9·3 ± 0·2)	0·6		

4.8. ^{9}Be *Target*

Compound system, ^{13}C ; binding energy, 26·287 MeV

The reaction channels open in ^{3}He bombardment of ^{9}Be are illustrated in figure 43.

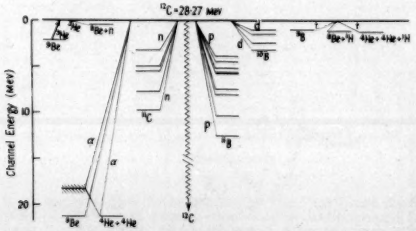


Figure 43. Schematic representation of reaction channels that are open when ^{9}Be is bombarded by 2 MeV ^{3}He ions. From Almqvist (1958).

^{9}Be(^{3}He, n)^{11}C $\qquad Q_m = 7{\cdot}566$ MeV

Detection of neutrons from this reaction at $E_{^3He} = 720$ kev has been reported (Moak, Good and Kunz 1954). Positron activity corresponding to the decay of ^{11}C produced by ^{3}He bombardment at about 20 MeV has been studied by Pool (1952) and Kundu *et al.* (1953).

As shown in figure 14 (Almqvist *et al.* 1956, 1960) the 90° yield of neutrons increases monotonically with incident energy in the range from 1·0 to 2·6 MeV—figure 31 compares the neutron spectrum from this reaction, as measured by

Stewart, van de Zwan and Almqvist (1959), with the corresponding spectrum from the ^{9}Be(^{3}He, p)^{11}B reaction measured by Almqvist *et al.* (1959). Neutron groups are observed in this spectrum corresponding to states in ^{11}C at excitations of 1·90, 4·23, 4·77, 6·46, 6·87, 7·39 and 8·12 MeV.

^{9}Be(^{3}He, p)^{11}B $Q_m = 10{\cdot}330$ MeV

Measurements on the energy spectra of protons from this reaction at $E_{^3He} = 900$ keV have been reported by Almqvist, Allen and Bigham (1955). Excitation functions for proton groups to low lying states in ^{11}B show a characteristic rise with incident energy below $E_{^3He} \sim 1{\cdot}8$ MeV followed by a relatively constant value thereafter. Figure 14 is characteristic of the data on the range from 1·0 to 2·6 MeV (Almqvist *et al.* 1956). Angular distributions of proton groups to the ground and first three states of ^{11}B have been measured at incident energies of 1·62,

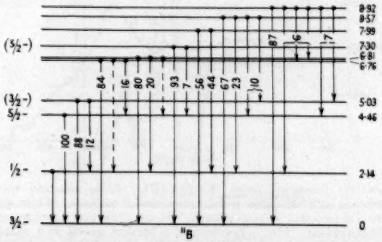


Figure 44. De-excitation branching data obtained from study of the ^{9}Be(^{3}He, pγ)^{11}B reaction gamma radiation. From Ferguson *et al.* (1958, 1959).

2·03 and 2·51 MeV (Almqvist *et al.* 1956); at 2·0 MeV (Holmgren, Bullock and Kunz 1955, 1956 a); at 4·5 MeV (Wolicki *et al.* 1959 a) and at 6·05 MeV (Sweetman 1958). Excitation functions for the proton groups to ground and first excited states have been measured for $1{\cdot}8 < E_{^3He} = 5{\cdot}2$ MeV at laboratory angles of 0°, 7°, 30°, 60°, 90°, 120° and 150°. Figures 15 and 16 illustrate some of the angular distributions measured for this reaction.

Detailed measurement on the gamma de-excitation of states in ^{11}B populated in this reaction have provided a number of the assignments shown in figure 44 (Ferguson *et al.* 1957, 1958 and 1960). These authors have compared the results obtained with the intermediate coupling shell model calculations of Kurath (1952, 1957), and have obtained excellent confirmation of these calculations.

^{9}Be(^{3}He, d)^{10}B $Q_m = 1{\cdot}092$ MeV

Observation of the gamma radiation associated with the de-excitation of the first excited state of ^{10}B populated in this reaction has been reported (Ferguson *et al.* 1957, 1960) using $E_{^3He} = 2{\cdot}1$ MeV.

The angular distribution of deuterons from this reaction to the ground and first four excited states of ^{10}B has been measured for $10° \lesssim \theta_d \leqslant 160°$ for $E_{^3He} = 25$ MeV

(Wegner and Hall 1958 a). All angular distributions show strong forward peaking with diffraction oscillations at forward angles ; beyond $\theta_{cl} = 90°$ the distributions are relatively flat.

The angular distributions corresponding to the clearly resolved levels have minima at 30° and 55° and maxima at 40° and ~70° in the centre-of-mass system. The angular distributions of groups to unresolved doublets show little diffraction structure since the levels formed are not all of the same parity (see § 2.3.3).

^{9}Be (^{3}He, α) ^{8}Be $\qquad Q_m = 18{\cdot}905$ MeV

Figure 45 shows the spectra of alpha particles from this reaction measured at 90° to the beam at $E_{^3He} = 0{\cdot}9$ MeV (Bigham 1954) ; measurements also have been

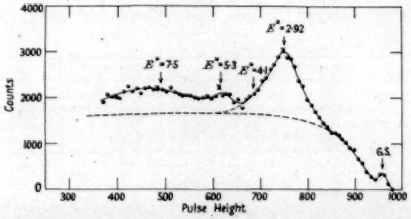


Figure 45. Alpha particle spectrum from the ^{9}Be (^{3}He, α) ^{8}Be reaction measured at 90° to an $E_{^3He} = 900$ kev beam with an argon filled proportional counter. Only alpha particle groups to the ground state and broad 2·92 MeV state in ^{8}Be are observed superposed on the ^{8}Be break-up alpha particle continuum. The suggestion of a further group at a pulse height of about 620 which would correspond to a state in ^{8}Be at approximately 5·3 MeV excitation is probably caused by particles hitting the centre wire of the counter. From Bigham (1954).

reported at 2·1 MeV by Almqvist *et al.* (1959). Only the weak ground state group of reaction alpha particles are clearly resolved from the continuum resulting from the break-up of ^{8}Be into two alpha particles. The group to the state in ^{8}Be at 2·92 MeV excitation is superposed on this continuum as shown ; its intensity is approximately 20 times that to the ground state.

^{9}Be (^{3}He, t) ^{9}B $\qquad Q_m = -1{\cdot}089$ MeV

Angular distributions of triton groups from this reaction have been reported by Wegner and Hall (1958 a) for $E_{^3He} = 21$ MeV ; all show forward peaking and diffraction structure. Figure 29 shows typical angular distributions measured.

4.9. ^{10}B *Target*

Compound system, ^{13}N ; binding energy, 21·645 MeV

^{10}B (^{3}He, n) ^{12}N $\qquad Q_m = 1{\cdot}248$ MeV

Neutron spectra and ground state angular distributions have been measured for $E_{^3He} = 2{\cdot}45$ MeV and 3·60 MeV (Ajzenberg-Selove, Bullock and Almqvist 1957 a, b). The ground state mass of ^{12}N was determined to be $12{\cdot}02255 \pm 0{\cdot}00007$ a.m.u. and excited states in ^{12}N were located at $1{\cdot}06 \pm 0{\cdot}08$, $1{\cdot}56 \pm 0{\cdot}08$, $(1{\cdot}97 \pm 0{\cdot}10)$, $2{\cdot}35 \pm 0{\cdot}08$,

3·18 ± 0·15, and 3·46 ± 0·15 in good agreement with the levels in the mirror nucleus ^{12}B (Ajzenberg-Selove and Lauritsen 1959, unpublished). For $E_{^3He} = 2{\cdot}45$ MeV the ground state neutron angular distribution was found to be isotropic; for $E_{^3He} = 3{\cdot}60$ MeV, the distribution was peaked forward with $I(0°)/I(140°) = 2{\cdot}9 \pm 0{\cdot}7$ in the centre-of-mass system. At $E_{^3He} = 2{\cdot}54$ MeV the ground state reaction cross section was measured to be $5{\cdot}2^{+2{\cdot}1}_{-1{\cdot}6}$ mbn; for $E_{^3He} = 3{\cdot}60$ MeV the 0° differential cross section is 0·73 ± 0·30 mbn sterad^{-1}. Figure 46 shows a typical neutron spectrum from this reaction.

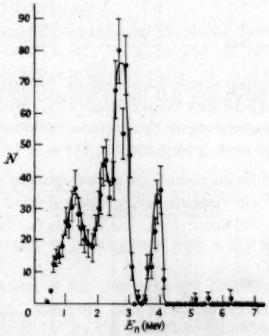


Figure 46. Neutron spectrum observed at 15° for 2·54 MeV ^{3}He bombardment of ^{10}B using photographic plate detection. This figure corresponds to the analysis of 1000 proton recoil tracks in the emulsion. The neutron groups in this spectrum correspond to the states of ^{12}N populated in the ^{10}B (^{3}He, n) ^{12}N reaction. From Ajzenberg-Selove, Bullock and Almqvist (1957).

^{10}B (^{3}He, p) ^{12}C $\qquad Q_m = 19{\cdot}701$ MeV

The 0° and 90° differential excitation function for this reaction has been measured for $1{\cdot}25 < E < 5{\cdot}0$ MeV (Schiffer *et al.* 1956 a, b). Table 13 lists the resonance structure found and the excitation energies of the corresponding states in the ^{13}N compound state. These resonances are superposed on a relatively smooth background yield which rises monotonically with energy for $E_{^3He} \lesssim 1{\cdot}8$ MeV and remains essentially constant thereafter.

Angular distributions of proton groups to the ground and 4·43 MeV states of ^{12}C have been measured at $E_{^3He} = 2{\cdot}00$ MeV (Johnston and Holmgren 1956 b), $E_{^3He} = 2{\cdot}17, 3{\cdot}17, 3{\cdot}62, 4{\cdot}06, 4{\cdot}42$ and 4·70 MeV (Schiffer *et al.* 1956 a, b) and at 6·05 MeV (Sweetman 1958). Deviations from symmetry relative to 90° in the centre-of-mass system are attributable to the resonances listed in table 13. At $E_{^3He} = 2{\cdot}00$ MeV the yield of ground state protons is only 10% of that to the 4·43 MeV state of ^{12}C.

Observation of proton groups to states in ^{12}C at excitations of $4{\cdot}43, 7{\cdot}77, 9{\cdot}61, 10{\cdot}75, 11{\cdot}83, 12{\cdot}76, 13{\cdot}31, 13{\cdot}97, 15{\cdot}10, 16{\cdot}04$ and 16·57 MeV have been reported by Bigham, Allen and Almqvist (1955) using $E_{^3He} = 900$ kev. Galonsky *et al.*

(1957) and Moak *et al.* (1958) observed the proton groups with $E_{^3He} = 1{\cdot}25$ MeV and at 15° to the beam as shown in figure 47 (Moak *et al.* 1958). Almqvist *et al.* (1959) report observation of proton groups to states in ^{12}C at excitations of 4·43, 7·76, 9·60, 10·76, 11·82, 12·78, 13·37, 14·03, 15·10, 16·10 and 16·64.

Table 13

E_{res} (MeV)	Γ_{res} (MeV)	E_{ex} (^{13}N) (MeV)
2·0	0·5	23·2
3·7	0·7	24·5
4·1	0·12	24·8
4·6	0·15	25·2

The only gamma-emitting states found below 15·5 MeV excitation in ^{12}C were those at 4·43, 12·78 and 15·10 MeV (Gove *et al.* 1957, Moak *et al.* 1958, Almqvist *et al.* 1959). From measurements on the gamma radiation de-exciting these states in coincidence with the proton group feeding them it has been shown (Almqvist

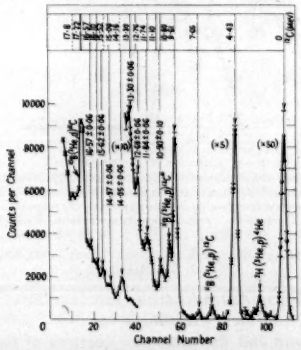


Figure 47. Proton spectrum from the ^{10}B (^3He, p) ^{12}C reaction showing groups corresponding to the excited states in ^{12}C at excitations below about 17 MeV. A corresponding ^{12}C level diagram is shown at the top of the figure. Contaminant reactions on ^2H and ^{11}B target contaminants are indicated. The energy of the incident ^3He beam was 1·25 MeV. From Moak *et al.* (1959).

et al. 1959) that in the case of the 15·10 MeV state, 97% of the de-excitation transitions are direct to the ground state while 3·1% are to the 4·43 MeV state. The total gamma radiation width of the state was found to be $(0{\cdot}75 \pm 0{\cdot}20)$ of the total width. The measured de-excitation branching ratio is in good accord with the predictions of an intermediate coupling shell model calculation (Kurath 1952, 1957).

Similar measurements (Almqvist *et al.* 1959) on the de-excitation of the 12·76 state show that $\Gamma_\gamma \sim 0{\cdot}02\,\Gamma_{total}$ and that less than 20% of the gamma transitions are via the 4·43 MeV state. The experimental data support the assignment 1^+, $T = 0$ as suggested on theoretical grounds (Kurath 1952, 1957).

Because the cross section for formation of the 4·43 MeV state exceeds that for the 7·77 MeV state by an order of magnitude (see table 5) the small but finite tail of the 4·43 proton group in the proton spectrum resulting from inelastic events in the detector crystals and extending under the group to the 7·77 MeV state precludes determination of a more stringent upper limit on Γ_γ for this state using this reaction than the value of 1 in 10^3 measured using a recoil technique by Eccles and Bodansky (1958). A measurement of this width is of particular interest in nucleogenesis calculations involving resonant radiative alpha particle capture by ^8Be via this state.

Moak *et al.* (1958) quote $\Gamma_\gamma/\Gamma = 50 \pm 25\%$, $3 \pm 1\%$ and $1{\cdot}4\%$ respectively for the 15·10, 12·78 and 7·77 MeV states.

The proton gamma angular correlation involving the 4·43 MeV state in ^{12}C has been measured at $E_{^3He} = 4{\cdot}5$ MeV. Anisotropies of 10%, 45% and 28% were observed when the proton detector was set at 10°, 45° and 90°, and 124° respectively. No symmetry relative to the nuclear recoil axis was observed (Johnston, Holmgren and Wolicki 1959).

$^{10}B(^3He, d)^{11}C$ $\qquad Q_m = 3{\cdot}206$ MeV

The deuteron group corresponding to formation of ^{11}C in the ground state is clearly evident in the charged particle spectrum from ^3He bombardment of ^{10}B shown in figure 47 (Moak *et al.* 1958). The deuteron group to the first excited state has also been observed in coincidence with a 2·0 MeV gamma ray (Almqvist *et al.* 1959).

The angular distribution of the ground state deuteron group was measured at $E_{^3He} = 2{\cdot}1$ MeV (Kuehner, Almqvist and Bromley 1958, unpublished) and show marked maxima at 17° and 75° in the centre-of-mass system, suggesting a direct interaction mechanism.

$^{10}B(^3He, \alpha)^9B$ $\qquad Q_m = 12{\cdot}134$ MeV

Proportional counter measurements on the spectrum from this reaction were carried out at $E_{^3He} = 1{\cdot}0$ MeV (Almqvist 1954). Two alpha particle groups were observed corresponding to formation of ^9B in its ground and first excited state at 2·33 MeV. No further groups with intensity greater than 20% of that to the ^9B ground state were found corresponding to excitations less than 6·5 MeV in ^9B.

Alpha particles from this reaction have been analysed magnetically for $E_{^3He} = 2{\cdot}0$, 2·5 and 3 MeV (Povh 1958, 1959). Alpha particle groups have been found leading to formation of ^9B in the ground, 2·33 MeV and 2·79 MeV states. The observed widths for the latter states are 80 and 300 KeV respectively; no evidence was found for a $1/2^+$ state in the region of $1{\cdot}7 - 1{\cdot}8$ MeV excitation nor were any well defined levels found at excitations between 4 and 7 MeV in ^9B.

Magnetic analysis of the alpha particle groups at $E_{^3He} = 5{\cdot}0$ MeV (Spencer and Phillips 1959) has located the ^9B excited states at $2{\cdot}39 \pm 0{\cdot}01$ and $2{\cdot}79 \pm 0{\cdot}05$ MeV. These states have widths of 83 ± 9 and 700 ± 200 KeV respectively. An alpha continuum was observed for all excitations greater than the sharp ground state suggesting that there may be states in ^9B below this sharp 'ground' state which are suppressed by the barrier for proton emission. These authors present evidence that the 2·39 MeV state in ^9B and the mirror state at 2·43 in ^9Be are very pure two-body states with a neutron or proton coupled to the 2^+, 2·9 MeV state of ^8Be.

Evidence is presented suggesting that this reaction may have a cusp in the alpha yield at the neutron threshold analogous to that observed in the $^{9}Be(pp')^{9}Be$ reaction, for example, at $1{\cdot}674 \pm 0{\cdot}005$ MeV.

4.10. ^{11}B *Target*

Compound system, ^{14}N ; binding energy, 20·724 MeV

$^{11}B(^{3}He, p)^{13}C$ $\qquad Q_m = 13{\cdot}184$ MeV

Excitation curves for the ground and first two excited state proton groups have been measured for $3{\cdot}00 < E_{^{3}He} < 5{\cdot}40$ MeV (Holmgren, Wolicki and Johnston 1957 a, 1959 a). The measured cross sections are roughly independent of the incident energy in this range. Differential cross sections of the seven highest energy proton groups were measured at $E_{^{3}He} = 4{\cdot}5$ and 5·4 MeV ; a number of these distributions show forward peaking ; however, no detailed analysis of these data such as that considered in § 2.3.2.3 has been attempted.

Measurements on the proton spectrum from this reaction at $E_{^{3}He} = 900$ kev have been reported by Bigham, Allen and Almqvist (1955). Eight proton groups were observed corresponding to levels in ^{13}C at excitations of 0, $3{\cdot}02 \pm 0{\cdot}08$, 3·68, $6{\cdot}96 \pm 0{\cdot}08$, $7{\cdot}67 \pm 0{\cdot}08$, $8{\cdot}96 \pm 0{\cdot}10$, $9{\cdot}98 \pm 0{\cdot}10$, $11{\cdot}2 \pm 0{\cdot}2$ and $11{\cdot}67 \pm 0{\cdot}10$ MeV.

Measurements at $E_{^{3}He} = 1{\cdot}25$ MeV (Galonsky *et al.* 1957, Moak *et al.* 1958) show proton groups to states in ^{13}C at $5{\cdot}51 \pm 0{\cdot}08$, $6{\cdot}10 \pm 0{\cdot}08$, $7{\cdot}58 \pm 0{\cdot}06$, $8{\cdot}85 \pm 0{\cdot}06$, $9{\cdot}54 \pm 0{\cdot}08$, $9{\cdot}92 \pm 0{\cdot}06$, $10{\cdot}9 \pm 0{\cdot}15$, $11{\cdot}1 \pm 0{\cdot}15$, $12{\cdot}08 \pm 0{\cdot}1$ and $12{\cdot}81 \pm 0{\cdot}1$ in addition to the well established states at 3·097, 3·684, 3·844 and 6·864 MeV which were used for calibration. The groups to states at 5·51 and 6·10 MeV are between 5 and 10% as intense as that to the ground state. Since these states have not been seen in other reactions leading to this nucleus, nor have the presumed mirror states in ^{13}N been identified, it is suggested that the states must have very narrow intrinsic widths.

The angular distributions of protons to low levels in ^{13}C have been measured at $E_{^{3}He} = 6{\cdot}05$ MeV (Sweetman 1958). All show characteristic forward peaking and have not as yet been subjected to any detailed theoretical analysis.

$^{11}B(^{3}He, d)^{12}C$ $\qquad Q_m = 10{\cdot}463$ MeV

The ground state excitation function has been measured for $3{\cdot}0 \leqslant E_{^{3}He} \leqslant 5{\cdot}40$ MeV (Holmgren, Wolicki and Johnston 1957 a, 1959 a) ; the deuteron yield increases linearly with energy by a factor of the order of 2 in this energy range. Angular distributions of the deuteron groups to the ground and first excited states of ^{12}C have been measured for $E_{^{3}He} = 4{\cdot}5$ and 5·4 MeV and show forward peaking strongly suggestive of a stripping mechanism. An attempt to fit these data is discussed in § 2.3.1.2 where the angular distributions are presented in figure 24.

The deuteron–gamma angular correlation involving de-excitation of the 4·43 MeV state in ^{12}C has been measured for $E_{^{3}He} = 4{\cdot}5$ MeV, and for deuteron observation angles of 7°, 30°, 60° and 124° in the laboratory (Wolicki *et al.* 1959). The measured anisotropies in these correlations are all in the range from 20–35% ; no evidence was found for symmetry of the correlation relative to the nuclear recoil axis in the reaction.

4.11. ^{12}C *Target*

Compound system, ^{15}O ; binding energy, 12·122 MeV

Figure 48 illustrates schematically the reactions resulting from ^{3}He bombardment of ^{12}C.

^{12}C(^{3}He, n)^{14}O $\qquad Q_m = -1{\cdot}167$ MeV

The neutron excitation function has been measured in the range

$$1{\cdot}30 \leqslant E_{^3\mathrm{He}} \leqslant 3{\cdot}75 \text{ MeV}$$

(Bromley *et al.* 1957). The threshold energy for the reaction was determined as 1449·6 ± 2·8 keV ; this corresponds to a mass defect for ^{14}O of 12·146 ± 0·0065 MeV and an end point for the ^{14}O positron decay of 1809·7 ± 7·8 KeV. This end point

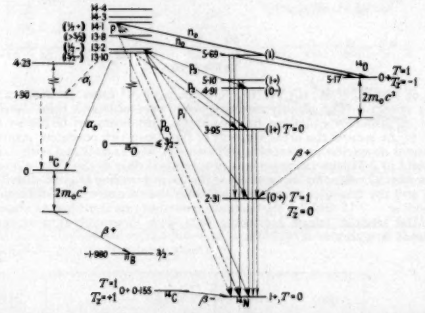


Figure 48. Schematic representation of the reactions induced by ^{3}He bombardment of a ^{12}C target at low incident energies. The ^{12}C (^{3}He, d) ^{13}N reaction has a calculated Q value of −3·550 MeV (table 2) and does not contribute to the reaction yield at low energies. From Bromley *et al.* (1957).

leads to a value for the Fierz constant in the Fermi beta interaction of 0·00 ± 0·02 (Gerhart 1958) and to a predicted lifetime of the muon of (2·26 ± 0·04) × 10^{-6} second (Feynmann and Gell-Mann 1958), to be compared with the experimental value of (2·22 ± 0·02) × 10^{-6} (Bell and Hincks 1951). Because of its importance in establishing the coupling constant in the Fermi interaction, more precise measurement of the ^{14}O positron end point would be desirable.

Angular distributions of the neutron group to the ^{14}O ground state have been measured at $E_{^3\mathrm{He}} = 1{\cdot}89$, 2·16, 2·40 and 2·51 MeV (Bromley *et al.* 1957). At the two lower energies the distributions are essentially isotropic while at the two higher single minima are found at centre-of-mass angles of 50° and 60° reflecting residual resonance structure in the yield curves. These distributions have been fitted to expansions of the form $W(\theta) = \sum_n a_n P_n(\cos\theta)$ for $n \leqslant 6$.

Assuming charge independence of nuclear forces, the cross sections for formation of the ^{14}O ground state and the analogue $T = 1$, $T_z = 0$, 2·313 MeV state in ^{14}N (figure 48) are predicted to be in the ratio 2 : 1 (Adair 1952) for equal neutron and

proton penetrabilities. Figure 49 (Bromley *et al.* 1957) shows this ratio as determined experimentally as a function of the incident ^{3}He energy. Again assuming charge independence, the angular distributions of neutrons and protons

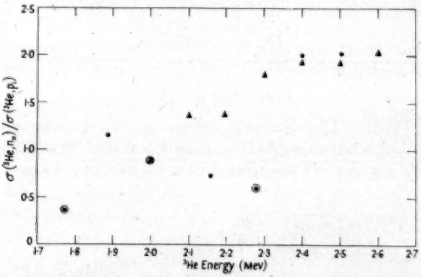


Figure 49. Ratio of the ^{13}C (^{3}He, n_0) ^{14}O to ^{13}C (^{3}He, p_1) ^{14}N* cross sections as a function of the incident ^{3}He energy. The doubly circled points were obtained from detailed measurements on the yield of 2·31 MeV radiation first with the ^{3}He beam on target so that the yield was proportional to the sum of the cross sections of the above two reactions, appropriately corrected for the measured de-excitation branching of states in ^{14}N, and second with the beam interrupted so that the yield of 2·31 MeV radiation was proportional only to the (^{3}He, n_0) cross section. The closed circles were obtained by integration of the corresponding angular distributions of neutrons and protons and the triangles were deduced from the measured 90° differential cross sections. The low points at ~2·15 and 2·3 MeV contain contributions from higher states in ^{14}N than were included in the analysis (since branching data were unavailable), and should therefore be ignored. From Bromley *et al.* (1957).

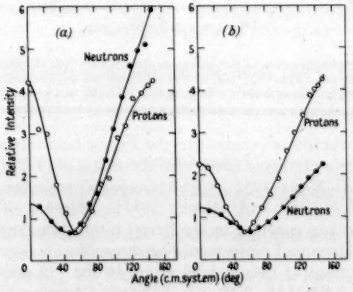


Figure 50. Comparison of the angular distributions of neutrons and protons from the ^{13}C (^{3}He, n) ^{14}O and ^{13}C (^{3}He, p) ^{14}N* (2·31 MeV) reactions. (*a*), $E_{^3He}$ = 2·51 MeV, (*b*) $E_{^3He}$ = 2·40 MeV. From Bromley *et al.* (1957).

to these analogue states are predicted to be identical. Figure 50 (Bromley *et al.* 1957) shows these corresponding angular distributions measured at 2·40 and 2·51 MeV. Clearly, both in figures 49 and 50, as the ^{3}He energy is raised above the neutron threshold, the experimental results support the charge independence

hypothesis. The presence of residual resonance structure in the excitation curve for this reaction (figure 51) implies that the compound states are sufficiently long-lived to allow mixing of the isobaric spin values in the compound system through the action of Coulomb forces.

The $^{12}C(^3He, n)^{14}O$ reaction constitutes a source of monoergic neutrons from 0 to about 6 MEV since no excited states in ^{14}O occur at excitations below 5·5 MEV (Ajzenberg and Franzen 1954). The reaction cross section rises rapidly with incident energy, as shown in figure 41, from 2·0 mbn at 1·9 MEV to 34 mbn at 2·5 MEV (Bromley *et al.* 1957).

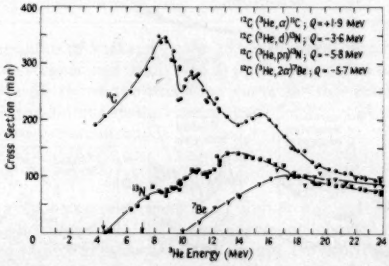


Figure 51. Excitation curves for the ^{12}C (^3He, n) ^{14}O and ^{12}C (^3He, α) ^{11}C and ^{12}C (^3He, 2α) ^7Be reactions as measured using stacked foil techniques and an incident ^3He energy of 21 MEV. From Cochran and Knight (1959, private communication).

$^{12}C(^3He, p)^{14}N$ $\quad Q_m = 4{\cdot}767$ MEV

Excitation curves for proton groups to the ground and first two excited states of ^{14}N have been measured in the energy ranges $1{\cdot}1 \leqslant E_{^3He} \leqslant 2{\cdot}9$ MEV (Bromley *et al.* 1957), $2{\cdot}0 \leqslant E_{^3He} \leqslant 5{\cdot}0$ MEV (Johnston *et al.* 1958).

The 90° differential excitation curve for the low energy range is shown in figure 52 (*a*). Figure 52 (*b*) extends these curves at this and other angles in the energy range $2{\cdot}0 \leqslant E_{^3He} \leqslant 5{\cdot}0$ MEV (Johnston *et al.* 1958). Resonance structure as listed in table 14 is observed in these curves. These resonances are superposed on a smoothly rising continuum yield. The reaction cross sections increase from about 0·03 mbn at $E_{^3He} = 1{\cdot}30$ MEV to about 30 mbn at $E_{^3He} = 2{\cdot}5$ MEV.

Angular distributions have been measured at several energies, $\leqslant 2{\cdot}66$ MEV (Bromley *et al.* 1957), 2·0 MEV (Johnston and Holmgren 1956 a), $\leqslant 5{\cdot}0$ MEV (Johnston *et al.* 1958) and at 6·05 MEV (Sweetman 1958). Figures 9, 25 and 12 in § 2.3.2.1 show typical distributions of groups to the first three states of ^{14}N. The angular distributions measured at lower energies (Bromley *et al.* 1957) have been fitted to Legendre polynomial expansions including sixth order polynomials.

Angular distributions measured at higher energies ($E_{^3He} = 21$ and 24 MEV) (Wegner and Hall 1958 a) show characteristic diffraction and forward peaking.

The branching ratio for the electromagnetic de-excitation of the 3·95 MEV state in ^{14}N has been measured; $(3{\cdot}7 \pm 0{\cdot}6)\%$ of the de-excitation transitions proceed

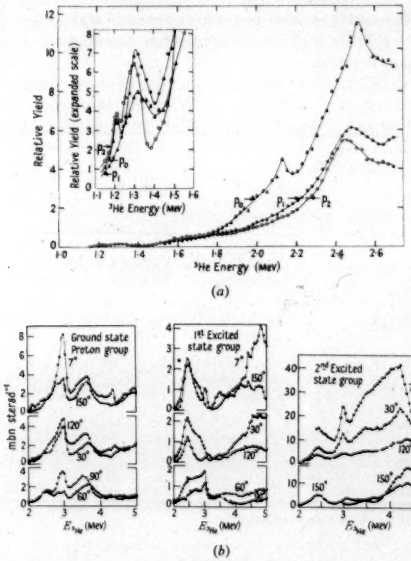


Figure 52. Excitation curves measured at 90°, for the ^{13}C (^3He, p) ^{14}N reaction groups leading to formation of ^{14}N in its ground and first two excited states. (*a*) Bromley *et al.* (1957), (*b*) Johnston *et al.* (1958).

Table 14. Resonance in the $^{13}C + ^3He$ Reaction

E_{res} (^3He) (MeV)	Resonant for groups	E_{ex} (^{15}O) (MeV)	J^π	Ref.
1·21	p_0, p_2	13·09	5/2⁻	(*a*)
1·3	p_0, p_1, p_2, p_3	13·2	—	(*a*)
2·15	p_0, n_0	13·84	5/2	(*a*)
2·52	p_0, p_1, p_2, p_3, n_0	14·14	1/2⁺	(*a*)
2·7	p_1, p_2, n_0	14·3		(*a*)
3·0	p_0, p_1, p_2	14·5		(*b*)
3·6	p_0, p_1, p_2	15·0		(*b*)
4·4	p_0, p_1, p_2	15·6		(*b*)
4·8	p_0, p_1, p_2	15·9		(*b*)
8·9	α_j, d_j	19·2		(*c*)
10·1	α_j, d_j	20·2		(*c*)
11·7	α_j, d_j	21·5		(*c*)
(11·9)	α_j, d_j	(21·6)		(*c*)

References : (*a*) Bromley *et al.* 1957, (*b*) Johnston *et al.* 1958, (*c*) Cochran and Knight 1958, 1959, private communication.

directly to the ground state while the remainder cascade via the 2·31 MeV level (Bromley *et al.* 1957, Gove *et al.* 1956). This is to be compared with the results of detailed calculations of Elliott (1956) and of Visscher and Ferrell (1955) giving $\sim 1\%$ and 3·9% respectively. Following the enunciation of the Morpurgo (1958 a, b) selection rule for M1 transitions in self-conjugate nuclei, it is clear that the case studied here presents a classic example of the operation of this selection rule which predicts a branching ratio $\sim 1\%$.

From measurements on the linear polarization of the 1·64 MeV cascade radiation (Litherland and Gove 1958) the parity of the 3·95 MeV state has been shown to be the same as that of the 2·31 MeV state hence positive (Bromley 1952).

$^{12}C(^3He, d)^{13}N$

Angular distributions of deuteron groups to low levels of ^{13}N have been measured for $E_{^3He} = 21$ MeV (Wegner and Hall 1958 a); all show diffraction structure (see figure 28). Figure 51 shows the excitation curve for this reaction as measured using a stacked foil technique in the energy range $5 \leqslant E_{^3He} \leqslant 24$ MeV (Cochran and Knight 1959, private communication).

$^{12}C(^3He, \alpha)^{11}C$ $\qquad Q_m = 1{\cdot}401$ MeV

Ground state reaction cross sections of 20 mbn at $E_{^3He} = 1{\cdot}78$ MeV and 50 mbn at $E_{^3He} = 20$ MeV (Bromley *et al.* 1957) and approximately 100 mbn in the range $5{\cdot}0 \leqslant E_{^3He} \leqslant 24$ MeV (figure 51) (Cochran and Knight 1959, private communication) have been determined from measurements on the 0·51 MeV annihilation radiation associated with the positron decay of the ^{11}C product nucleus.

Angular distributions of alpha particles to several of the low lying levels of ^{11}C have been measured for $E_{^3He} = 21$ MeV (Wegner and Hall 1958 a). All these distributions show characteristic diffraction structure.

4.12. ^{13}C *Target*

Compound system, ^{16}O; binding energy, 22·771 MeV

$^{13}C(^3He, p)^{15}N$ $\qquad Q_m = 7{\cdot}175$ MeV

0° and 90° excitation curves for this ground state reaction have been measured for $1{\cdot}2 \leqslant E_{^3He} \leqslant 4{\cdot}9$ MeV (Schiffer *et al.* 1956 a, b). Resonance structure superposed

Table 15

E_{res} (MeV)	Γ_{res} (MeV)	E_{ex} (MeV)
3·6	0·5	25·7
4·5	0·4	26·6

on a rising continuum is attributed to levels in ^{16}O as noted in table 15. Excitation curves have been measured in the range $1{\cdot}9 \leqslant E_{^3He} \leqslant 3{\cdot}8$ MeV (Illsley *et al.* 1957) at laboratory angles of 0°, 30°, 60°, 90°, 120° and 150°.

Angular distributions have been measured for $E_{^3He} = 2{\cdot}26$, 2·92, 3·62, 3·90 and 4·52 MeV (Schiffer *et al.* 1956, 1956 a) and at 2·0 and 4·5 MeV (Illsley *et al.* 1957).

All angular distributions show minima at about 90° in the centre-of-mass system, together with a forward maximum increasing in intensity with incident energy.

Studies on proton groups to states in ^{15}N at 5·3 and 6·3 MeV excitation have been reported (Johnston, Holmgren and Wolicki 1957).

^{13}C (^{3}He, α) ^{12}C $\quad Q_m = 15{\cdot}625$ MeV

Excitation curves have been measured for $1{\cdot}9 \leqslant E_{^3\mathrm{He}} \leqslant 5{\cdot}0$ MeV (Wolicki *et al.* 1956) at laboratory angles of 7°, 30°, 60°, 90° and 120°. No well resolved resonance structure was found although the 3·6 and 4·5 MeV resonances found in the ^{13}C (^{3}He, p) ^{15}N studies (above) do appear as inflections in the excitation curves; lesser evidence might be adduced for resonances at approximately 2·1, 2·7 and 3·1 MeV.

Angular distributions of the alpha particle groups to the ground and 4·43 MeV states of ^{12}C have been measured at 2·0 MeV (Holmgren 1956, 1957) and 4·5 MeV (Holmgren *et al.* 1956 c, 1957). These show very marked structure and are discussed in §2.3.1.1. Figure 19 presents the ground state angular distributions at these energies.

The alpha–gamma angular correlation involving the de-excitation of the 4·43 MeV level populated in this reaction has been measured for $E_{^3\mathrm{He}} = 4{\cdot}5$ MeV, and for alpha particle detection angles of 10°, 40°, 90° and 120° in the laboratory are shown in figure 21 (Holmgren, Wolicki and Johnston 1959 a). These correlations show strong structure with 5 : 1 peak to valley structure $\theta_\alpha = 10°$ and 1·7 : 1 for $\theta_\alpha = 90°$. The correlations are approximately of the form $1 + A\cos^2\theta_\gamma + B\cos^4\theta_\gamma$ but show no symmetry relative to the nuclear recoil axes and are not amenable to simple analysis (see §3.3). Gamma radiation from the de-excitation of the 15·09 MeV level in ^{12}C has been observed for $E_{^3\mathrm{He}} = 2{\cdot}1$ MeV (Almqvist *et al.* 1959).

4.13. ^{14}N *Target*

Compound system, ^{17}F; binding energy, 15·826 MeV

Figure 53 is a schematic representation of the ^{3}He induced reactions on ^{14}N.

^{14}N (^{3}He, n) ^{16}F $\quad Q_m = -1{\cdot}129$ MeV

A search for the threshold for this reaction carried out for $E_{^3\mathrm{He}} \leqslant 2{\cdot}0$ MeV (Butler 1956) yielded negative results. This implies a mass defect of at least 16·1 MeV for ^{16}F and instability of the ground state against proton emission by at least 1·3 MeV can be found in this reaction. The threshold value is calculated from the nuclear mass data (table 2) to be at 1·21 MeV.

^{14}N (^{3}He, p) ^{16}O $\quad Q_m = 15{\cdot}231$ MeV

Excitation curves for proton groups to states in ^{16}O at excitations below 11 MeV have been measured for $2{\cdot}1 \leqslant E_{^3\mathrm{He}} \leqslant 2{\cdot}62$ MeV. These curves show no resonant structure in this region (Bromley *et al.* 1957 a, 1959 e).

Angular distributions of protons leading to the states at ~6·0, ~7·0, 8·87, ~10 MeV have been found to be essentially isotropic (Bromley *et al.* 1957 a, 1959 e).

Magnetic analysis of the proton groups from this reaction (Bromley *et al.* 1959 e) at $E_{^3\mathrm{He}} = 2{\cdot}70$ MeV has located an unnatural parity ($\pi = (-1)^{J+1}$) doublet of levels

at excitations of 10·94 and 11·09 MeV in ^{16}O, in addition to the level at 8·87 MeV (Wilkinson, Toppel and Alburger 1956).

Detailed gamma–gamma angular correlation measurements on the de-excitation gamma radiation from these unnatural parity states has permitted unambiguous spin and parity assignments (Kuehner *et al.* 1959) using the equal magnetic substate

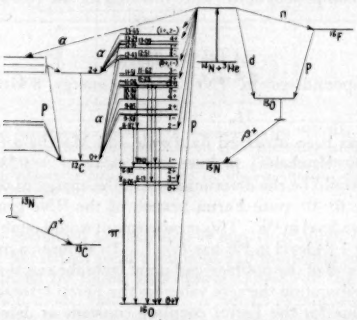


Figure 53. Schematic representation of the reactions induced by ^3He bombardment of a ^{14}N target at low incident energies. Only states of unnatural parity (i.e. $\pi = (-1)^{J+1}$) at excitations greater than about 7 MeV in ^{16}O de-excite by emission of gamma radiation. From Bromley *et al.* (1959 e).

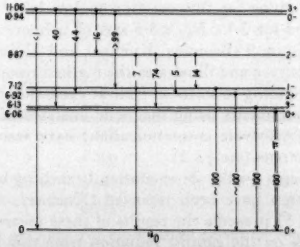


Figure 54. De-excitation branching of states in ^{16}O populated in the ^{14}N (^3He, p) ^{16}O reaction. From Bromley *et al.* (1959 e).

population hypothesis discussed in §2.2.3. These assignments, together with the experimentally measured de-excitation branching ratios, are shown in figure 54. These data, together with those on the transition multipole mixtures determined from the gamma–gamma correlation measurements, have been compared with intermediate coupling shell model calculations of Elliott and Flowers (1957). In general, very satisfactory agreements obtained. From this comparison it is possible to establish the exchange character of the ^{16}O potential (Bromley *et al.* 1957 a,

1959 e, Kuehner *et al.* 1959); that found is very similar to the Serber rather than the Rosenfeld mixture commonly assumed in these calculations.

From the experimentally determined value of $\Gamma_\alpha/\Gamma_\gamma \lesssim 0{\cdot}20$ for the 10·94 MeV state de-excitation, and from the systematics of alpha and gamma de-excitation widths (Wilkinson 1958) an upper limit of 2×10^{-9} has been established on the intensity of the opposite parity component of the wave function for the 10·94 MeV state (Bromley *et al.* 1959 e).

4.14. ^{16}O *Target*

Compound system, ^{19}Ne ; binding energy, 8·416 MeV

^{16}O(^{3}He, n)^{18}Ne $\qquad Q_m = -2{\cdot}967$ MeV

This reaction has been observed by Towle and Macefield (1959, private communication from R. Batchelor) to have a Q value of $-3{\cdot}235 \pm 0{\cdot}020$ MeV. Of particular interest would be the determination of the energy of the gamma radiation associated with the 0^+–0^+ pure Fermi branch of the ^{18}Ne ground state positron decay to the analogue level in ^{18}F. This measurement would establish unambiguously which of the 1·08 or 1·04 level in ^{18}F has $T = 1$. In addition a precise measurement of the ^{18}Ne lifetime and of the positron end point from the reaction energies involved is of interest in corroborating the zero value for the Fermi interaction Fierz constant and the current value for the Fermi coupling constant as determined from other similar decays (e.g. ^{14}O, ^{26}Al, ^{34}Cl etc.) (Gerhart 1958). ^{3}He energies in excess of 3·8 MeV are required for convenient study of this reaction.

^{16}O(^{3}He, p)^{18}F $\qquad Q_m = 2{\cdot}053$ MeV

The excitation functions for this reaction to low lying states ($E_{ex} \leqslant 3$ MeV) in ^{18}F have been measured for $2{\cdot}1 \leqslant E_{^3He} \leqslant 5{\cdot}5$ MeV (Kuehner, Almqvist and Bromley 1958 a, b, 1959 b). Figure 9 (Bromley, Kuehner and Almqvist 1959 b) shows the lower section of these curves and illustrates the marked resonance structure present in this reaction corresponding to states in ^{19}Ne at excitation energies between about 11 and 15 MeV. Measurements using magnetic analyses of the emergent protons (Povh and Barnes 1959, private communication) have resolved further structure in these excitation functions (see § 3.2).

Detailed measurements on the de-excitation branching of the low levels of ^{18}F populated in this reaction have been reported (Kuehner, Almqvist and Bromley 1958 a, 1960). Figure 55 presents the results of these measurements.

Early measurements on the gamma radiation from this reaction using 2·0 MeV ^{3}He ions (Butler, Holmgren and Kunz 1956) located a doublet of levels at approximately 1 MeV excitation. Later magnetic spectrometric measurements (Kuehner, Almqvist and Bromley 1958 b) have established the existence of the four levels shown at an excitation of about 1 MeV. Subsequent magnetic spectrographic measurements at higher incident energy (Hinds and Middleton 1959 a) have established the existence of the levels listed in table 16 (see also figure 30). These levels have also been examined by Rao *et al.* (1959).

Lifetime measurements on the de-excitation of the 0·94 MeV state have shown that $\tau \geqslant 10^{-13}$ sec (Naggiar *et al.* 1956) and $\tau \lesssim 5 \times 10^{-9}$ sec (Kuehner, Almqvist and Bromley 1958 a). The parity of this state has been established as equal to that of

the ground state from linear polarization measurements on the de-excitation gamma radiation (Litherland and Gove 1958). Measurements on the lifetime of the state

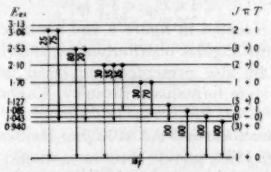


Figure 55. De-excitation branching for states in ^{18}F populated in the ^{16}O (^{3}He, p) ^{18}F reaction. The level assignments listed are based on measurements of the branching ratios, angular distributions and angular correlations for the radiations involved. The bracketed assignments are very probable ones but cannot be completely fixed by the experimental data currently available. From Kuehner, Almqvist and Bromley (1960).

at 1·112 MeV for de-excitation to that at 0·94 MeV (Eccleshall 1959, private communication from K. W. Allen) yield a value of 110 mμ sec in excellent agreement with the value predicted by Elliott and Flowers (1958) on the basis of an intermediate coupling shell model calculation. A detailed comparison of the data on

Table 16. Energy Levels in ^{18}F ; E_{ex} = Excitations in MeV†

E_{ex}	J	π	T	E_{ex}	J	π	T	E_{ex}	J	π	T
0	1	+	0	3·352				4·840			
0·940	3	+	0	3·727				4·965			
1·042	0	−	0	3·790				5·292			
1·087	0	+	1	3·841				5·500			
1·129	5	+	0	4·116				5·603	1		0
1·699	1	+	0	4·227				5·666			
2·105	2	+	0	4·358				5·785			
2·525	3	+	0	4·400				6·093			
3·063	2	+	1	4·649				6·137			
3·131				4·741				6·232			

The level assignments are consistent with the measured gamma ray branching (Kuehner, Almqvist and Bromley 1958 a) and the angular distributions observed in ^{14}N (α, γ) ^{18}F (Almqvist, Bromley and Kuehner 1958).

the low levels of ^{18}F with these predictions and those of Redlich (1958) from a similar calculation has been carried out (Kuehner, Almqvist and Bromley 1958 a, 1960 ; Bromley, Kuehner and Almqvist 1958) ; in general, quite satisfactory agreement has been found supporting the validity of these calculations.

† *Note added in proof.* Recent measurements on the decay of ^{18}Ne by J. W. Butler (1960, private communication) together with measurements on the ^{16}O (^{3}He, p$_3$) ^{18}F and ^{16}O (^{2}H, p$_0$) ^{17}O proton angular distributions at ~5·5 MeV by Hinds and Middleton and Jaffe (to be published), establish the assignment 0^+, $I = 1$ to the 1·043 MeV state in ^{18}F leaving 0^-, $T = 0$ as the most probable assignments to the 1·085 MeV state.

$^{16}O(^3He, \alpha)^{15}O$ $\quad Q_m = 4{\cdot}976$ MeV

The ground state reaction shows strong resonance behaviour for

$$2{\cdot}1 \leqslant E_{^3He} \leqslant 3{\cdot}2 \text{ MeV}.$$

These data have been presented in figure 9 and are discussed in detail in §2.2.1 where measurements of the angular distribution expansion coefficients as functions of the incident energy are also presented and fitted using standard interfering resonance, compound system formalism (Bromley, Kuehner and Almqvist 1958 a, b, 1960).

Measurements have been carried out at higher incident energies, of more than 3·5 MeV (Povh and Barnes 1959, private communication) ; magnetic analysis of the reaction alpha particles has been used to identify the level doublet at 5·174 and 5·233 MeV excitation in ^{15}O. This doublet, corresponding to the 5·276 and 5·305 MeV states in ^{15}N, is also reported by Hinds and Middleton (1959 b) from magnetic analysis of the $^{16}O(^3He, \alpha)^{15}O$ alpha particles at $E_{^3He} = 5{\cdot}700 \pm 0{\cdot}010$ and $5{\cdot}905 \pm 0{\cdot}010$ MeV.

4.15. ^{19}F *Target*

Compound system, ^{23}Na ; binding energy, 18·655 MeV

$^{19}F(^3He, ^3He)^{19}F$

Coulomb excitation measurements using both ^3He and ^4He ions in the energy range from 1·0 – 2·5 MeV (Bromley, Kuehner and Almqvist 1959, 1959 a) have been carried out to test the applicability of the yield ratio technique for transition multipolarity determination (Bjerregaard and Huus 1954). Both the E1 excitation of the $1/2^-$, 110 kev state and the E2 excitation of the $5/2^+$, 197 kev state were studied. Thick target excitation curves for both the E1 and E2 ^3He induced transitions are in good agreement with the theoretical predictions (Alder *et al.* 1956).

$^{19}F(^3He, \alpha)^{18}F$ $\quad Q_m = 10{\cdot}171$ MeV

Magnetic analysis of the reaction alpha particles has corroborated the existence of the levels found in the $^{16}O(^3He, p)^{18}F$ reaction. Angular distributions of these alpha particle groups have been measured at several energies ; these show patterns strongly suggestive of simple stripping mechanism (Hinds and Middleton 1959 a).

4.16. ^{20}Ne *Target*

Compound system, ^{23}Mg ; binding energy, 13·319 MeV

$^{20}Ne(^3He, p)^{22}Na$

Using 3·5 MeV ^3He ions and a neon gas target of normal isotopic abundances Temmer and Heydenburg (1958) have examined the proton spectra leading to low levels in ^{22}Na. Levels existed at excitations of 0, (0·593, 0·666), 0·89, 1·54, (1·94, 1·99, 2·22), 2·58, (2·98, 3·07) and 3·53 MeV, where the bracketed levels were not clearly resolved. Proton–gamma and gamma–gamma coincidence measurements using a fast–slow system (figure 32) have permitted observation of the 0·073 MeV transition from the 666 to 593 kev states. Evidence is advanced for the assignments

0^+, $T = 1$, and 1^+, $T = 0$ to these levels respectively. Delayed coincidence measurements have been used to determine the half-lives of these states in ^{22}Na as 0·014 and 0·266 μsec respectively.

4.17. ^{24}Mg *Target*

Compound system, ^{27}Si ; binding energy, 13·370 MeV

^{24}Mg (^{3}He, α) ^{23}Mg $\quad Q_m = 4{\cdot}008$ MeV

Magnetic analysis of the alpha groups from this reaction at

$$E_{^3\mathrm{He}} = 5{\cdot}905 \pm 0{\cdot}010 \text{ MeV}$$

locates states in ^{23}Mg at excitations of 0·451, 2·048, 2·356, 2·712, 2·768, 2·904, 3·792, 3·856, 3·968 and 4·353 MeV. With the exception of the last value where an error of 15 kev is quoted, a 10 kev error has been assigned to these excitation values (Hinds and Middleton 1959 b). These authors compare these excitations with those in the mirror system ^{23}Na and find excellent correspondence.

4.18. ^{28}Si *Target*

Compound system, ^{31}S ; binding energy, 12·437 MeV

^{28}Si (^{3}He, α) ^{27}Si $\quad Q_m = 3{\cdot}374$ MeV

Magnetic analysis of the alpha groups from this reaction at $E_{^3\mathrm{He}} = 5{\cdot}700$ and 5·891 locates states in ^{27}Si at excitations of 0·782, 0·958, 2·165, 2·651, 2·866, 2·908, 3·540, 3·800 and perhaps 4·13 MeV. An error of 10 kev has been assigned to each of these except the latter two where errors of 15 and 20 kev apply (Hinds and Middleton 1959 b).

4.19. ^{40}Ca *Target*

Compound system, ^{43}Ti ; binding energy, 9·941 MeV

^{40}Ca (^{3}He, d) ^{41}Sc

Wegner and Hall (1959, private communication) have examined the differential cross section for this reaction at $E_{^3\mathrm{He}} = 23{\cdot}65 \pm 0{\cdot}15$ MeV leading to the ground and first four excited states of ^{41}Sc. The angular distributions to the ground and first two excited states show characteristic diffraction oscillations such as those shown in figure 28 (§ 3.3.3) for the ^{12}C (^{3}He, d) ^{13}N reaction, with that to the first excited state out of phase with the other two. It should be noted that in a stripping formalism $l_p = 3$ is required for the ground state of ^{41}Sc while $l_p = 1$ suffices for the first excited state. In contrast to these results, the angular distributions to the third and fourth excited states show a monotonic decrease with increasing angle. These results suggest that these states have much more complex wave functions than do the lower three states.

4.20. *Higher A Targets*

Using activation techniques and incident ^{3}He ions from cyclotrons, excitation functions for many reactions leading to radioactive final nuclei have been studied. Summaries of the measurements carried out using the beam from the Ohio State

cyclotron have been given by Pool (1952) and Kundu *et al.* (1953) and using the beam from the Birmingham cyclotron by Fremlin (1952).

An extensive series of measurements using the Los Alamos variable energy cyclotron is at present in progress. Cochran and Knight are utilizing activation techniques while Wegner and Hall are using a particle identification $dE/dx \times E$ system to study the individual reaction kinematics in greater detail. As yet only very preliminary reports of this work are available (Cochran and Knight 1958, (1959, private communication), Wegner and Hall 1958 a, 1959).

Results of Coulomb excitation measurements on ^{23}Na, ^{47}Ti and ^{55}Mn are presented in § 3.4 (Bromley, Kuehner and Almqvist 1959, 1959 a).

§ 5. Suggested Experiments

5.1. *General Comments*

As is clear from the data presented in §§ 3 and 4 of this report, the field of study of ^{3}He induced reactions involving target masses of more than 20 a.m.u. is essentially untouched as yet. Even in the lighter nuclei, with few exceptions, only survey measurements have been reported. This implies a need for much more extensive measurements at low incident energies and more particularly in the range from 5–20 Mev which is available from some existing cyclotrons and in part at least from tandem Van de Graaff generators.

In this section we shall attempt to enumerate a few of the fields in which we believe further study might be particularly profitable ; the number of such problems yet unexamined is legion.

5.2. *Nuclear Spectroscopic Investigations*

As noted throughout this report, the high Q values associated with ^{3}He reactions make possible the detailed study of the de-excitation of states in final nuclei at high excitations even at low incident energies. For energies significantly below the Coulomb barrier, however, the reaction yields are strongly inhibited and this has in considerable measure limited the scope of published research to date. With the recent and impending availability of significantly higher energy ^{3}He accelerators it becomes feasible to consider extending measurements such as those reported here throughout the periodic table. In particular, it appears reasonable to expect that application of the equal magnetic substate hypothesis discussed in §§ 2.2.3 and 3.3 to the study of gamma–gamma de-excitation cascades will yield a very large amount of nuclear spectroscopic data as will measurements carried out in the special geometries described in § 3.3.

The small amount of such detailed spectroscopic evidence to date has limited the detailed testing of nuclear models to the regions of light nuclei or of closed nuclear shells where the level structure is particularly simple.

It should be emphasized, however, that the experimental problems involved change with increasing ^{3}He energy. In particular at higher excitations in light final nuclides and even at low excitations in heavier ones, the level density is such that the relatively low resolution detectors (e.g. crystal scintillators) which have been widely used to date are inadequate to separate individual reaction groups. Under these conditions, and in addition, because of the larger number of open

reaction channels at higher energies, it becomes necessary to use magnetic spectrometric and crystal or other techniques in conjunction to obtain the required energy resolution and particle identification.

5.3. *Angular Distribution and Correlation Measurements*

As discussed in § 2, the experimental information at present available does not lead to a sufficiently clear understanding of ^{3}He induced reaction mechanisms to permit their use as spectroscopic tools for the determination of level parameters as has been the case for deuteron reactions. This ambiguity may in large measure reflect the fact that essentially all of these measurements are at relatively low energies where Coulomb effects are particularly important.

As found by Wegner and Hall (1958 a) at energies around 20–30 MeV, the reaction data are remarkably insensitive to the detailed nuclear structure apart from the nuclear shape. It would appear most profitable therefore to examine angular distributions from (^{3}He, α) and (^{3}He, d) reactions for $\sim 8 \leqslant E_{^3\mathrm{He}} \leqslant 15$ MeV in more detail. Having elucidated the mechanism involved, these reactions should serve as fruitful sources of spectroscopic data. Moreover, from a determination of the form factors involved in these angular distributions, detailed information will be obtainable on the wave function of the ^{3}He ion itself and on the corresponding 3-body problems.

By comparing the results of ^{3}He and ^{2}H (and where available ^{3}H) induced reactions involving single nucleon transfer, it will be possible to isolate and identify Coulomb effects with greater precision than has hitherto been possible.

Despite their additional inherent complexity, the (^{3}He, p) and (^{3}He, n) reaction angular distributions appear capable of providing relevant data on nuclear structure as evidenced by the work of El Nadi and El Khishin discussed in § 2.3.2.1 for example.

Apart from the pioneering work of the Naval Research Laboratory (U.S.A.) group on the study of particle–gamma angular correlations, nothing has been done in this field. This provides a very sensitive and powerful probe for determining details of the reaction mechanisms and of distortion effects arising from interaction of the incident and outgoing ions with the target and residual nuclei respectively. Such distortions have been neglected entirely in the Born approximation calculations reported herein. They have been considered in considerable detail for the analogous cases of ^{2}H induced reactions by Satchler in numerous publications to which the reader is referred and by Austern (1959) in his review treatment of direct interaction mechanisms.

5.4. *Isobaric Multiplet Studies*

Use of ^{3}He ions facilitates detailed measurements in isobaric multiplet states as has been noted previously. To date only measurements in the $A = 5$, 6, 7, 9, 11, 12, 14, 23 and 27 systems have been reported. Extension of these measurements in detail and scope will provide additional evidence on the validity of the charge independence hypothesis and on the extent to which isobaric spin may be considered as a good quantum number throughout the periodic table.

In the case of odd–odd systems in particular, it becomes possible to determine the positron end point energies in $0^+ \rightarrow 0^+$, pure Fermi, decays with greater precision

than is possible by direct positron measurements, through determination of the corresponding (^{3}He, n) threshold values. Such information is particularly relevant to the determination of the coupling and Fierz constants for the Fermi interaction.

In the case of ^{18}Ne, for example, determination of the ft value for the decay from measurements on the ^{16}O(^{3}He, n)^{18}Ne reaction is of particular interest. In addition, examination of the gamma radiation coincident with this decay would identify which of the 1·04 or 1·08 Mev states in ^{18}F is the $T = 1$ analogue of the ^{18}Ne ground state.† Equivalent measurements are of considerable interest throughout the periodic table but require ^{3}He energies of 5 Mev or over.

5.5. *Neutron Measurements*

From an examination of table 2 it is clear that by using any one of the neutron detection techniques in current use, as well as neutron–gamma coincidence detectors, on (^{3}He, n) reactions, it will be possible to obtain information on the level structure of a large number of proton-rich nuclides, many of which have not been studied at all as yet. The reader is referred to Bromley and Almqvist (1959) for a survey of the relevant techniques.

5.6. *Lifetime Measurements*

To provide more stringent tests of any nuclear model, measurements of excited state lifetimes (i.e. the state widths for electromagnetic, and in some cases particle de-excitation), in addition to the more usual measurements of ground state beta decay lifetimes, are important. Such measurements are of particular interest in the cases of new proton-rich nuclides because of the information which they provide relevant to beta decay theories. Several of the nuclear recoil techniques (Devons 1958, Severiens and Hanna 1956, Burde and Cohen 1956) are particularly applicable to such measurements on ^{3}He induced reactions.

5.7. *Shell Effects*

Since the (^{3}He, n) reaction, for example, provides a means of adding two protons simultaneously to the target nucleus, it provides a convenient probe for the study of the nuclear shell structure. It would be of interest, for example, to examine the relative (^{3}He, n) cross sections in situations where the two added protons go into the middle of an unfilled proton shell, into the first or last two places in the shell, or when one proton goes into the last place and the second into the next shell. For such measurements it should be possible to gain considerable direct insight into the pairing energies and shell energies involved. Such a series of measurements on the argon isotopes, for example, would provide this evidence in the region of the $Z = 20$ proton shell. It should be noted that the shell model formalism predicts a large cross section enhancement for the situation where the captured nucleons complete a particular shell (Lane and Elliott 1958).

5.8. *Coulomb Excitation Measurements*

As demonstrated in § 3.4, observation of the relative Coulomb excitation yields with ^{3}He and ^{4}He ions of corresponding energy permits a determination of the transition multipolarity without the attendant complications involved in absolute

† See footnote to table 16.

cross section determinations. With the energies available to date it has not been feasible to carry out such measurements in the very heavy nuclei with large permanent deformations. The large deformations, hence moments of inertia, imply that the level spacings are relatively small, hence the higher order levels are accessible to helium ion induced excitation rather than the case in intermediate mass nuclei where heavier ions are required to give significant excitation of the corresponding levels. Use of helium ions has a marked advantage over heavier ions in these ratio measurements, despite the lower intrinsic cross sections, because the differences between the ratio values, characteristic of the various transition multipoles, decrease rapidly with increasing mass of the incident ion. With helium ions, relatively inaccurate ratio determinations ($\pm 5\%$) suffice to determine the multipolarity involved whereas with heavier ions, such as ^{16}O and ^{18}O, for example, measurements to approximately 1% are required.

Identification of the various 1^- and 3^- states in even–even heavy nuclei corresponding to pear-shaped deformations and static nuclear octopole moments, if these latter exist (Bohr and Mottelson 1958), will be of crucial importance to the further understanding of the foundations of the nuclear collective model (Bohr and Mottelson 1955).

5.9. *Nucleon Cluster Phenomena*

As noted previously in §2.3.1.1, detailed measurements on all of the (^{3}He, p), (^{3}He, d), (^{3}He, ^{3}He) and (^{3}He, α) reactions on a given target nucleus provide a determination of the relative reaction amplitudes. This in turn is intimately related to the existence of nucleon clusters in the nuclear surface.

Since in essentially all of the many-body nuclear theories considered to date, all but binary nucleon–nucleon interactions have been neglected, data bearing on the nucleon-clustering, if present, can lead to important modification of the model wave functions and certainly to the predicted system binding energy.

A systematic series of such detailed measurements carried out on targets selected at intervals through a nuclear shell would provide interesting information on the correlations, if any, between nuclear deformation and surface nucleon clustering.

§ 6 Conclusions

In this report we have attempted to survey the field of ^{3}He induced reactions in the light of measurements published to date and the further measurements which these suggest.

Almost without exception, the comments which have been made here relevant to studies with ^{3}He ions apply with equal force to those with ^{3}H ions.

In situations where low energy (i.e. 1–3 MeV) accelerators are available we would strongly suggest consideration of a programme of study of ^{3}He or ^{3}H induced reactions. The additional complexity involved in gas recovery or recirculation is relatively trivial ; with the decreasing cost of ^{3}He gas supplies even this becomes unnecessary. The absence of an associated toxicity health hazard is an advantage of ^{3}He over ^{3}H ; however, in the latter case the system requirements to eliminate this hazard are minor ones.

Our conclusion is that there is tremendous scope for fruitful experimental work with even a very low energy accelerator equipped to use either of these ions.

ACKNOWLEDGMENTS

We are indebted to Drs. J. A. Kuehner and E. Vogt for critical reading of this review and to Mrs. Alison Macfarlane for typing the manuscript.

REFERENCES

ADAIR, R. K., 1952, *Phys. Rev.*, **87**, 1041.
AJZENBERG, F., and FRANZEN, W., 1954, *Phys. Rev.*, **94**, 409.
AJZENBERG-SELOVE, F., BULLOCK, M. L., and ALMQVIST, E., 1957 a, *Bull. Amer. Phys. Soc.*, Ser. II, **2**, 28; 1957 b, *Phys. Rev.*, **108**, 1284.
AJZENBERG-SELOVE, F., and LAURITSEN, T., 1959, *Nuclear Physics*, **11**, 1.
ALDER, K., BOHR, A., HUUS, T., MOTTELSON, B., and WINTHER, A., 1956, *Rev. Mod. Phys.*, **28**, 432.
ALDRICH, L. T., and NIER, A. C., 1946, *Phys. Rev.*, **70**, 983; 1948, *Ibid.*, **74**, 1590.
ALLEN, K. W., and ALMQVIST, E., 1953, *Rev. Sci. Instrum.*, **24**, 70.
ALLEN, K. W., ALMQVIST, E., and BIGHAM, C. B., 1955, *Phys. Rev.*, **99**, 631.
ALLISON, S. K., and WARSHAW, S. D., 1953, *Rev. Mod. Phys.*, **25**, 779.
ALLISON, S. K., and NORBECK, E., 1956, *Rev. Sci. Instrum.*, **27**, 285.
ALLISON, S. K., 1958, *Rev. Mod. Phys.*, **30**, 1137.
ALLRED, J. C., 1951, *Phys. Rev.*, **84**, 695.
ALLRED, J. C., ARMSTRONG, A. H., HUDSON, A. M., POTTER, R. M., ROBINSON, E. S., ROSEN, L., and STOVALL, E. J., 1952, *Phys. Rev.*, **88**, 425.
ALMQVIST, E., ALLEN, K. W., DEWAN, J. T., PEPPER, T. P., and SANDERS, J. H., 1950, *Phys. Rev.*, **79**, 1034.
ALMQVIST, E., ALLEN, K. W., DEWAN, J. T., and PEPPER, T. P., 1951, *Phys. Rev.*, **83**, 202; 1953, *Ibid.*, **91**, 1022.
ALMQVIST, E., 1954, *Ph.D. Thesis*, University of Liverpool.
ALMQVIST, E., ALLEN, K. W., and BIGHAM, C. B., 1955, *Phys. Rev.*, **99**, 631.
ALMQVIST, E., BROMLEY, D. A., GOVE, H. E., LITHERLAND, A. E., and PAUL, E. B., 1956, *Bull. Amer. Phys. Soc.*, Ser. II, **1**, 195.
ALMQVIST, E., 1958, *Proceedings* 1958 *Accelerator Conference, High Voltage Engineering Corp., Burlington, Mass., U.S.A.*
ALMQVIST, E., BROMLEY, D. A., and KUEHNER, J. A., 1958, *Bull. Amer. Phys. Soc.*, Ser. II, **3**, 27.
ALMQVIST, E., BROMLEY, D. A., FERGUSON, A. J., GOVE, H. E., and LITHERLAND, A. E., 1959, *Phys. Rev.*, **114**, 1040.
ALMQVIST, E., BROMLEY, D. A., CLARKE, R. L., GOVE, H. E., and PAUL, E. B., 1960, *Canad. J. Phys.*, in the press.
ALVAREZ, L. W., and CORNOG, R., 1939, *Phys. Rev.*, **56**, 379, 613.
VON ARDENNE, M., 1956, *Atomkern-Energie*, **4**, 121, translated 1957 by E. L. Pode, *A.E.R.E. Lib/Trans.*, 758.
ARNOLD, W. R., 1952, *Rev. Sci. Instrum.*, **23**, 97.
ARNOLD, W. R., PHILLIPS, J. A., SAWYER, G. A., STOVALL, E. J., and TUCK, J. L., 1954, *Phys. Rev.*, **93**, 483.
ASHENBRENNER, F. A., 1954, *Phys. Rev.*, **95**, 600.
ATKINS, K. R., and LOVEJOY, D. R., 1954, *Canad. J. Phys.*, **32**, 702.
AUSTERN, N., 1960, *Nuclear Spectroscopy, Part A*, Ed. F. Ajzenberg-Selove (New York: Academic Press).
BAINBRIDGE, K. T., 1953, *Experimental Nuclear Physics*, Vol. 1, Ed. E. Segrè (New York: Wiley).
BAIR, J. K., and MAIENSCHEIN, F. C., 1951, *Rev. Sci. Instrum.*, **22**, 343.
BAKER, G. A., HOLLOWAY, J. T., KING, L. D. P., and SCHREIBER, R. E., 1948, *U.S.A.E.C. Report* AECD-2189.
BALL, J. B., FAIRHALL, A. W., and HALPERN, I., 1958, *Bull. Amer. Phys. Soc.*, Ser. II, **3**, 322.
BARKAS, W. H., 1939, *Phys. Rev.*, **56**, 1242.
BARTHOLOMEW, G. A., and KINSEY, B. B., 1952, *Phys. Rev.*, **86**, 605; 1953, *Canad. J. Phys.*, **31**, 537.

BARTHOLOMEW, G. A., KNOWLES, W., and LEE-WHITING, G., 1960, *Rep. Progr. Phys.*, **23**, 453 (London : Physical Society).
BATCHELOR, R., 1955, *Proc. Phys. Soc.* A, **68**, 452.
BATCHELOR, R., and TOWLE, J. H., 1959, *U.K.A.E.A. Report* NR/P–12/58.
BATCHELOR, R., and MORRISON, G. C., 1960, *Fast Neutron Physics*, Ed. J. B. Marion and J. L. Fowler (New York : Interscience).
BATCHELOR, R., FERGUSON, A. J., GOVE, H. E., and LITHERLAND, A. E., 1960, *Nuclear Physics*, in the press.
BAY, Z., 1950, *Phys. Rev.*, **77**, 419.
BAYLY, A. J., and WARD, A. G., 1948, *Canad. J. Res.* A, **26**, 69.
BELING, J. K., NEWTON, J. O., and ROSE, B., 1952, *Phys. Rev.*, **86**, 797 ; **87**, 670.
BELL, P. R., 1955, *Beta and Gamma Spectroscopy*, Ed. K. Siegbahn (Amsterdam : North-Holland).
BELL, R. E., and PETCH, H. E., 1949, *Phys. Rev.*, **76**, 1409.
BELL, R. E., GRAHAM, R. L., and PETCH, H. E., 1952, *Canad. J. Phys.*, **30**, 35.
BELL, R. E., 1955, *Beta and Gamma Spectroscopy*, Ed. K. Siegbahn (Amsterdam : North-Holland).
BELL, W. E., and HINCKS, E. P., 1951, *Phys. Rev.*, **84**, 1243.
BENNETT, E. F., 1958, *Bull. Amer. Phys. Soc.*, Ser. II, **3**, 26.
BERLIN, T. H., and OWEN, G. E., 1958, *Nuclear Physics*, **5**, 669.
BIEDENHARN, L. C., BLATT, J. M., and ROSE, M. E., 1952, *Rev. Mod. Phys.*, **24**, 249.
BIEDENHARN, L., 1960, *Nuclear Spectroscopy, Part A*, Ed. F. Ajzenberg-Selove (New York : Academic Press).
BIGHAM, C. B., 1954, *Ph.D. Thesis*, University of Liverpool.
BIGHAM, C. B., ALLEN, K. W., and ALMQVIST, E., 1955, *Phys. Rev.*, **99**, 631.
BIRKS, J. B., 1953, *Scintillation Counters* (New York : McGraw-Hill).
BITTNER, J. W., 1954, *Rev. Sci. Instrum.*, **25**, 1058.
BJERREGAARD, J. H., and HUUS, T., 1954, *Phys. Rev.*, **94**, 204.
BLAIR, J. M., HINTZ, N. M., and VAN PATTER, D. M., 1954, *Phys. Rev.*, **96**, 1023.
BLAIR, J. S., and HENLEY, E. M., 1958, *Phys. Rev.*, **112**, 2029.
BLAIR, J. S., 1960, *Phys. Rev.*, in the press.
BLATT, J. M., and WEISSKOPF, V. K., 1952, *Theoretical Nuclear Physics* (New York : Wiley).
BLAUGRUND, A. E., 1958, *Proc. Rehovoth Conf. Nuclear Structure*, p. 542, Ed. H. J. Lipkin (Amsterdam : North-Holland).
BOHR, A., and MOTTELSON, B. R., 1953, *K. Danske Vidensk. Selsk., Mat-Fys. Medd.*, **27**, No. 16 ; 1957, *Nuclear Physics*, **4**, 529.
BONNER, T. W., CONNER, J. R., and LILLIE, A. B., 1952, *Phys. Rev.*, **88**, 473.
BONNER, T. W., and COOK, C. F., 1954, *Phys. Rev.*, **96**, 122.
BOOTH, D. L., HILL, R. S., PRICE, F. V., and ROAF, D., 1957, *Proc. Phys. Soc.* A, **70**, 863.
BORKOWSKI, C. J., 1959, *Proceedings of the Tri-partite Conference on Nuclear Instrumentation* (Chalk River, April 1959).
BOWRING, R. W., 1957, *U.K.A.E.A. Report, A.E.R.E.* GP/R 2058.
BOWRING, R. W., and DAVIES, R. H., 1958, *U.K.A.E.A. Report, A.E.R.E.* GP/R 2058.
BOYER, K., NORTHRUP, J. A., and STOKES, R. H., 1957, *Bull. Amer. Phys. Soc.*, Ser. II, **2**, 207.
BRADBURY, J. N., and STEWART, L., 1958, *Bull. Amer. Phys. Soc.*, Ser. II, **3**, 417.
BROLLEY, J. E., FOWLER, J. L., and STOVALL, E. J., 1951, *Phys. Rev.*, **82**, 502.
BROLLEY, J. E., HALL, J. H., and ROSEN, L., 1957, *Bull. Amer. Phys. Soc.*, Ser. II, **2**, 207.
BROMLEY, D. A., 1952, *Phys. Rev.*, **88**, 565.
BROMLEY, D. A., and BRUNER, J. A., 1954, *U.S.A.E.C. Document* NYO-3823. Reissued as *Chalk River Report* CRP-884, 1959.
BROMLEY, D. A., GOVE, H. E., LITHERLAND, A. E., PAUL, E. B., and ALMQVIST, E., 1956, *Bull. Amer. Phys. Soc.*, Ser. II, **1**, 195.
BROMLEY, D. A., ALMQVIST, E., GOVE, H. E., LITHERLAND, A. E., PAUL, E. B., and FERGUSON, A. J., 1957, *Phys. Rev.*, **105**, 957.
BROMLEY, D. A., FERGUSON, A. J., GOVE, H. E., LITHERLAND, A. E., and ALMQVIST, E., 1957 a, *Bull. Amer. Phys. Soc.*, Ser. II, **2**, 51.

BROMLEY, D. A., GOVE, H. E., LITHERLAND, A. E., PAUL, E. B., and ALMQVIST, E., 1957 b, *Canad. J. Phys.*, **35**, 1042.
BROMLEY, D. A., 1958, *Proc. of the Rehovoth Conference*, ed. by H. K. Lipkin (Amsterdam : North-Holland).
BROMLEY, D. A., and RUTLEDGE, A. R., 1958, *Chalk River Report* CRP-789.
BROMLEY, D. A., KUEHNER, J. A., and ALMQVIST, E., 1958, *Bull. Amer. Phys. Soc.*, Ser. II, **3**, 27 ; 1958 a, *Ibid.*, Ser. II, **3**, 199 ; 1958 b, *Compte Rendu, C.I.R.N.*, Paris.
BROMLEY, D. A., and ALMQVIST, E., 1959, *Chalk River Report* CRP-881.
BROMLEY, D. A., KUEHNER, J. A., and ALMQVIST, E., 1959, *Bull. Amer. Phys. Soc.*, Ser. II, **4**, 8 ; 1959 a, *Phys. Rev.*, **115**, 586 ; 1959 b, *Nuclear Physics*, **13**, 1.
BROMLEY, D. A., FERGUSON, A. J., GOVE, H. E., KUEHNER, J. A., LITHERLAND, A. E., ALMQVIST, E., and BATCHELOR, R., 1959 c, *Canad. J. Phys.*, **37**, 1574.
BROMLEY, D. A., GOVE, H. E., KUEHNER, J. A., LITHERLAND, A. E., and ALMQVIST, E., 1959 e, *Phys. Rev.*, **114**, 758.
BROOKS, F. D., 1959, *Nuclear Instr. and Methods*, **4**, 151.
BROWN, R. J. S., FREIER, G., HOLMGREN, H. D., STRATTON, W. R., and YARNELL, J. L., 1952, *Phys. Rev.*, **88**, 253.
BROWN, R. J. S., FAMULARO, K. F., HOLMGREN, H. D., and VAN PATTER, D. M., 1953, *Phys. Rev.*, **91**, 438A.
BROWN, R. J. S., FAMULARO, K. F., HOLMGREN, H. D., RANKIN, R. L., and STRATTON, W. R., 1954, *Phys. Rev.*, **96**, 80.
BROWNE, C. P., and BUECHNER, W. W., 1956, *Rev. Sci. Instrum.*, **27**, 899.
BUCHELNIKOVA, N. S., 1958, *Adv. Phys. Sci., Moscow*, **65**, No. 3.
BUECHNER, W. W., STRAIT, E. N., STERGIOPOULOS, C. G., and SPERDUTO, A., 1948, *Phys. Rev.*, **74**, 1569.
BUNBURY, D. ST. P., DEVONS, S., MANNING, G., and TOWLE, J. H., 1956, *Proc. Phys. Soc.* A, **69**, 165.
BURDE, J., and COHEN, S. G., 1956, *Phys. Rev.*, **104**, 1093.
BUTLER, J. W., 1956, *Bull. Amer. Phys. Soc.*, Ser. II, **1**, 94.
BUTLER, J. W., HOLMGREN, H. D., and KUNZ, W. E., 1956, *Bull. Amer. Phys. Soc.*, Ser. II, **1**, 29.
BUTLER, S. T., 1951, *Proc. Roy. Soc.* A, **208**, 559.
BUTLER, S. T., and SYMONDS, J. L., 1951, *Phys. Rev.*, **83**, 858.
BUTLER, S. T., and SALPETER, E. E., 1952, *Phys. Rev.*, **88**, 133.
BUTLER, S. T., 1957, *Nuclear Stripping Reactions* (New York : Wiley).
BUTLER, S. T., AUSTERN, N., and PEARSON, C., 1958, *Phys. Rev.*, **112**, 1227.
CALDWELL, D. O., 1952, *Phys. Rev.*, **88**, 131.
CAMERON, A. G. W., 1958, *Bull. Amer. Phys. Soc.*, Ser. II, **3**, 227.
CHIBA, R., CONZETT, H. E., MORINAGA, H., MUTSURO, N., SHODA, K., and KIMURA, M., 1958, *Bull. Amer. Phys. Soc.*, Ser. II, **3**, 419.
CLAUSNITZER, G., FLEISCHMANN, R., and SCHOPPER, H., 1956, *Z. Phys.*, **144**, 336.
CLUSIUS, K., and DICKEL, G., 1938, *Naturwiss.*, **26**, 546 ; 1939, *Z. phys. Chem.*, **44**, 397.
COCHRAN, D. R. F., and KNIGHT, J. D., 1958, *Bull. Amer. Phys. Soc.*, Ser. II, **3**, 323 ; 1959, *Ibid.*, **4**, 219.
COLLINS, L. E., and RIVIERE, R. C., 1959, *Nuclear Instr. and Methods*, **4**, 121.
CONNER, J. R., BONNER, T. W., and SMITH, J. P., 1952, *Phys. Rev.*, **88**, 468.
COOK, C. W., FOWLER, W. A., LAURITSEN, C. C., and LAURITSEN, T., 1957, *Phys. Rev.*, **107**, 508.
COON, J. H., 1949, *Phys. Rev.*, **75**, 1355.
CRANBERG, L., 1955, *Int. Conf. on Peaceful Uses of Atomic Energy*, Vol. 4, 577.
CRANBERG, L., BEAUCHAMP, R. K., and LEVIN, J. S., 1957, *Rev. Sci. Instrum.*, **28**, 89.
CROSS, W. G., 1951, *Rev. Sci. Instrum.*, **22**, 717.
CURRAN, S. C., 1953, *Luminescence and the Scintillation Counter* (New York : Academic Press).
CURRAN, S. C., and CRAGGS, J. D., 1949, *Counting Tubes* (New York : Academic Press).
DANFORTH, J. L., 1958, *Canad. Elect. Eng.*, **2**, 18.
DAS GUPTA, N. N., and GHOSH, S. K., 1946, *Rev. Mod. Phys.*, **18**, 227.
DAUNT, J. G., 1952, *Advanc. Phys.*, **1**, 209.

DEMERS, P., 1959, *Ionographie* (Montreal : University Press).
DEMEUR, M., 1954, *Physica*, **20**, 1189.
DEVONS, S., MANNING, G., and BUNBURY, D. ST. P., 1955, *Proc. Phys. Soc.* A, **68**, 18.
DEVONS, S., MANNING, G., and TOWLE, J. H., 1956, *Proc. Phys. Soc.* A, **69**, 173.
DEVONS, S., and GOLDFARB, L. J. B., 1957, *Handbuch der Physik*, Vol. XLII, p. 460, Ed. S. Flugge (Berlin : Springer).
DEVONS, S., 1958, *Proc. Rehovoth Conf. Nucl. Structure*, Ed. H. J. Lipkin, p. 547 (Amsterdam : North-Holland).
DONAVEN, T. W., KUNDU, D. N., LONG, J. K., and POOL, M. L., 1952, *Phys. Rev.*, **88**, 158.
DROZDOV, S. I., 1955, *Soviet Physics, JETP*, **1**, 588, 591 ; *Ibid.*, **7**, 889.
DUKEL'SKU, V. M., AFROSIMOV, V. V., and FEDORENKO, N. V., 1956, *Zh. eksper. teor. Fiz.*, **30**, 792.
DUMOND, J. W. M., 1955, *Beta and Gamma Spectroscopy*, Chap. IV, Ed. K. Siegbahn (Amsterdam : North-Holland).
DUNNING, K. L., BUTLER, J. W., and BONDELID, R. O., 1956, *Bull. Amer. Phys. Soc.*, Ser. II, **1**, 328 ; 1958, *Phys. Rev.*, **110**, 1076.
ECCLES, S. F., and BODANSKY, D., 1958, *Bull. Amer. Phys. Soc.*, Ser. II, **3**, 188.
EFFAT, K. E. A., and FREMLIN, J. H., 1955, *J. Sci. Instrum.*, **32**, 363.
EL NADI, M., 1957, *Proc. Phys. Soc.* A, **70**, 62.
EL NADI, M., and EL KHISHIN, M., 1959, *Proc. Phys. Soc.*, **73**, 705.
ELLIOTT, J. P., 1956, *Phil. Mag.*, **6**, 503.
ELLIOTT, J. P., and FLOWERS, B. H., 1955, *Proc. Roy. Soc.* A, **229**, 536; 1958, *Ibid.*, A, **242**, 57.
ELLIOTT, L. G., and BELL, R. E., 1948, *Phys. Rev.*, **74**, 1869.
ENDT, P. M., and BRAAMS, C. M., 1957, *Rev. Mod. Phys.*, **29**, 683.
ENGE, H. A., 1958, *Rev. Sci. Instrum.*, **29**, 885.
EVANS, J. E., KINSEY, B. B., and WHITMORE, B. G., 1959, *Canad. J. Phys.*, **37**, 396.
FAGG, L. W., and HANNA, S. S., 1959, *Rev. Mod. Phys.*, **31**, 711.
FARMER, B. J., and CLASS, C. M., 1959, *Bull. Amer. Phys. Soc.*, Ser. II, **4**, 95.
FELD, B., 1953, *Experimental Nuclear Physics*, Vol. 2, Ed. E. Segrè (New York: Wiley).
FERENTZ, M. I., and ROSENZWEIG, N., 1955, Argonne National Laboratory, Report ANL–5324, *Table of Coefficients*.
FERGUSON, A. J., and RUTLEDGE, A. R., 1957, *Chalk River Report* CRP-615.
FERGUSON, A. J., GOVE, H. E., LITHERLAND, A. E., ALMQVIST, E., and BROMLEY, D. A., 1957, *Bull. Amer. Phys. Soc.*, Ser. I, **2**, 51 ; 1958, *Phys. Rev. Lett.*, **1**, 414 ; 1960, *Phys. Rev.*, in the press.
FEYNMANN, R., and GELL-MANN, M., 1958, *Phys. Rev.*, **109**, 193.
FISCHER, G., 1957, *Bull. Amer. Phys. Soc.*, Ser. II, **2**, 28.
FLOWERS, B. H., 1950, *Proc. Roy. Soc.* A, **204**, 503.
FOWLER, W. A., 1958, *Astrophys. J.*, **127**, 551.
FREIER, G., and HOLMGREN, H., 1954, *Phys. Rev.*, **93**, 825.
FREMLIN, J. H., 1952, *Proc. Phys. Soc.* A, **65**, 762 ; 1953, *Phil. Mag.*, **44**, 141.
FRETTER, W. B., 1955, *Introduction to Experimental Physics* (New York : Prentice Hall).
FULBRIGHT, H. W., 1958, *Handbuch der Physik*, Vol. 45, Ed. S. Flugge and E. Creutz (Berlin : Springer).
FULTON, T., and OWEN, G. E., 1957, *Phys. Rev.*, **108**, 789.
FURRY, W. H., and JONES, R. C., 1946, *Rev. Mod. Phys.*, **18**, 156.
GALONSKY, A., MOAK, C. D., TRAUGHBER, R. L., and JONES, C. M., 1957, *Bull. Amer. Phys. Soc.*, Ser. II, **2**, 51.
GEER, E., HOLMGREN, H. D., JOHNSTON, R. L., and WOLICKI, E. A., 1956, *Bull. Amer. Phys. Soc.*, Ser. II, **1**, 211.
GEER-ILLSLEY, E., HOLMGREN, H. D., JOHNSTON, R. L., and WOLICKI, E. A., 1957, *Phys. Rev.*, **107**, 538.
GERHART, J. B., 1958, *Phys. Rev.*, **109**, 897.
GERHOLM, T. R., 1956, *Arkiv för Fysik*, **11**, 55.
GOLDFARB, L. J. B., 1958, *Nuclear Reactions*, Ed. P. M. Endt and M. Demeur (Amsterdam : North-Holland).

Good, W. M., Kunz, W. E., and Moak, C. D., 1951, *Phys. Rev.*, **83**, 845 ; 1954, *Ibid.*, **94**, 87.
Goodman, C. D., and Need, J. L., 1958, *Phys. Rev.*, **110**, 676.
Goodwin, L. K., 1953, *Rev. Sci. Instrum.*, **24**, 635.
Gove, H. E., Litherland, A. E., Almqvist, E., and Bromley, D. A., 1956, *Bull. Amer. Phys. Soc.*, Ser. II, **1**, 196 ; 1956 a, *Phys. Rev.*, **103**, 835.
Gove, H. E., Litherland, A. E., Almqvist, E., Bromley, D. A., and Ferguson, A. J., 1957, *Bull. Amer. Phys. Soc.*, Ser. II, **2**, 51.
Gove, H. E., and Litherland, A. E., 1958, *Bull. Amer. Phys. Soc.*, Ser. II, **3**, 200.
Gove, H. E., and Rutledge, A. R., 1958, *Chalk River Report* CRP-755.
Gove, H. E., 1958, *Nuclear Reactions*, Chap. VI, Ed. P. M. Endt and M. Demeur (Amsterdam : North-Holland).
Gove, H. E., and Litherland, A. E., 1960, *Nuclear Spectroscopy, Part A*, Ed. F. Ajzenberg-Selove (New York : Academic Press).
Grodzins, L., 1958, *Proceedings of the Geneva Conference on the Peaceful Uses of Atomic Energy*, **14**, 351 (New York : United Nations).
Gugelot, P. C., and Phillips, P. R., 1956, *Phys. Rev.*, **101**, 1614.
Gugelot, P. C., and Rickey, M., 1956, *Phys. Rev.*, **101**, 1613.
Hale, D. H., and Present, R. D., 1955, *Phys. Rev.*, **99**, 1366.
Hansen, A. O., and McKibben, J. L., 1947, *Phys. Rev.*, **72**, 673.
Hatton, J., and Preston, G., 1949, *Nature, Lond.*, **164**, 143.
Heath, R. L., 1957, *U.S.A.E.C. Research and Development Report* IDO-16408 (TID-4500).
Henkel, R. L., Perry, J. E., and Smith, R. K., 1955, *Phys. Rev.*, **99**, 1050.
Henley, E. M., 1958, *Conference on Photonuclear Reactions* (N.B.S., Washington, U.S.A.).
Heydenburg, N. P., and Temmer, G. M., 1956, *Annual Review of Nuclear Science*, **6**, 77.
Hinds, S., and Middleton, R., 1959 a, *Proc. Phys. Soc.*, **73**, 721 ; 1959 b, *Proc. Phys. Soc.*, **73**, 727.
Hintz, N. M., Blair, J. M., and Van Patter, D. M., 1954, *Phys. Rev.*, **93**, 924A.
Hoffman, C. J., Edeskuty, F. J., and Hemmel, E. F., 1956, *J. Chem. Phys.*, **24**, 124.
Holmgren, H. D., Bullock, L., and Kunz, W. E., 1955, *Phys. Rev.*, **100**, 436.
Holmgren, H. D., Kunz, W. E., and Bullock, L., 1955 a, *Bull. Amer. Phys. Soc.*, Ser. I, **30**, 26.
Holmgren, H. D., 1956, *Bull. Amer. Phys. Soc.*, Ser. II, **1**, 21.
Holmgren, H. D., Bullock, L., and Kunz, W. E., 1956 a, *Phys. Rev.*, **104**, 1446.
Holmgren, H. D., Geer, E., Johnston, R. L., and Wolicki, E. A., 1956 b, *Bull. Amer. Phys. Soc.*, Ser. II, **1**, 325.
Holmgren, H. D., Johnston, R. L., Wolicki, E. A., and Geer, E., 1956 c, *Bull. Amer. Phys. Soc.*, Ser. II, **1**, 211.
Holmgren, H. D., 1957, *Phys. Rev.*, **106**, 100.
Holmgren, H. D., Geer, E., Johnston, R. L., and Wolicki, E. A., 1957, *Phys. Rev.*, **106**, 102.
Holmgren, H. D., Wolicki, E. A., and Johnston, R. L., 1957 a, *Bull. Amer. Phys. Soc.*, Ser. II, **2**, 181.
Holmgren, H. D., and Johnston, R. L., 1958, *Bull. Amer. Phys. Soc., Ser.* II, **3**, 26 ; 1959, *Phys. Rev.*, **113**, 1556.
Holmgren, H. D., Wolicki, E. A., and Johnston, R. L., 1959, *Bull. Amer. Phys. Soc.*, Ser. II, **4**, 232 ; 1959 a, *Phys. Rev.*, **114**, 1281.
Holøien, E., and Midtdal, J., 1955, *Proc. Phys. Soc.* A, **68**, 815 ; 1958 a, *Nucl. Instrum. and Methods*, **3**, 57.
Hoogenboom, A. M., 1958, *Ph.D. Thesis* (University of Utrecht).
Howland, P. R., Scofield, N. E., and Taylor, R. A., 1955, *U.S. Navy Research and Development Laboratory Report* USNRDL-TR-654, NS-081-001.
Huby, R., 1952, *Proc. Roy. Soc.* A, **215**, 385.
Illsley, E., Holmgren, H. D., Johnston, R. L., and Wolicki, E. A., 1957, *Phys. Rev.*, **107**, 538.
Inopin, E. V., 1957, *Soviet Physics, JETP*, **4**, 764.
Irving, J., 1951, *Phil. Mag.*, **42**, 338.
Ise, J., Pyle, R. V., Hicks, D. A., and Main, R. M., 1954, *Rev. Sci. Instrum.*, **25**, 437.

JARMIE, N., and SEAGRAVE, J. D., *et al.*, 1957, *Los Alamos Report* 2014.
JARVIS, R. G., and ROAF, D., 1953 a, *Proc. Roy. Soc.*, **218**, 432 ; 1953 b, *Proc. Phys. Soc.* A, **66**, 310.
JOHANSSON, S. A. E., 1950, *Nature, Lond.*, **166**, 794.
JOHNSON, C. H., JUDISH, J. P., and SNYDER, C. W., 1957, *Rev. Sci. Instrum.*, **28**, 942.
JOHNSON, J. L., and PORTER, W. S., 1955, *Phys. Rev.*, **99**, 630A.
JOHNSTON, R. L., and HOLMGREN, H. D., 1956 a, *Bull. Amer. Phys. Soc.*, Ser. II, **1**, 21 ; 1956 b, *Ibid.*, Ser. II, **1**, 134.
JOHNSTON, R. L., WOLICKI, E. A., GEER, E., and HOLMGREN, H. D., 1956, *Bull. Amer. Phys. Soc.*, Ser. II, **1**, 197.
JOHNSTON, R. L., HOLMGREN, H. D., and WOLICKI, E. A., 1957, *Bull. Amer. Phys. Soc.*, Ser. II, **2**, 181.
JOHNSTON, R. L., HOLMGREN, H. D., WOLICKI, E. A., and GEER-ILLSLEY, E., 1958, *Phys. Rev.*, **109**, 884.
JOHNSTON, R. L., HOLMGREN, H. D., and WOLICKI, E. A., 1959, *Bull. Amer. Phys. Soc.*, Ser. II, **4**, 233.
JONES, W. M., 1955, *Phys. Rev.*, **100**, 124.
JORGENSEN, T., KUYATT, C., and SAUTTER, C. A., 1958, *Bull. Amer. Phys. Soc.*, Ser. II, **3**, 230.
JUDD, D., 1950, *Rev. Sci. Instrum.*, **21**, 213.
KISTIAKOWSKY-FISCHER, V., 1957, *Bull. Amer. Phys. Soc.*, Ser. II, **2**, 28.
KLIUCHAREV, A. P., ESEL'SON, B. N., and VAL'TER, A. K., 1956, *Dokl. Akad. Nauk, S.S.S.R.*, **109**, 475.
KNOWLES, W., 1957, *Chalk River Report* GPI–42.
KOMAROV, V. V., NEUDACHIN, V. G., POPOVA, A. M., and TEPLOV, I. B., 1959, *Soviet Physics, JETP*, **35**, (8), 679.
KORFF, S. A., 1946, *Electron and Nuclear Counters* (New York : Van Nostrand).
KRAVTSOV, V. A., 1958, *Uspekhi Fiz. Nauk*, **65** (3), 451.
KUEHNER, J. A., ALMQVIST, E., and BROMLEY, D. A., 1958 a, *Bull. Amer. Phys. Soc.*, Ser. II, **3**, 27 ; 1958 b, *Phys. Rev. Lett.*, **1**, 260 ; 1960, to be published.
KUEHNER, J. A., LITHERLAND, A. E., ALMQVIST, E., BROMLEY, D. A., and GOVE, H. E., 1959, *Phys. Rev.*, **114**, 775.
KUNDU, D. N., and POOL, M. L., 1953, *Phys. Rev.*, **91**, 417.
KUNDU, D. N., DONAVEN, T. W., POOL, M. L., and LONG, J. K., 1953, *Phys. Rev.*, **89**, 1200.
KUNZ, W. E., 1955, *Phys. Rev.*, **97**, 456 ; 1953, *Ibid.*, **91**, 462A.
KUNZ, W. E., BUTLER, J. W., and HOLMGREN, H. D., 1955 a, *Bull. Amer. Phys. Soc.*, Ser. I, **30**, 25 ; 1955 b, *Phys. Rev.*, **100**, 1252A.
KUNZ, W. E., MOAK, C. D., and GOOD, W. M., 1953, *Phys. Rev.*, **91**, 676 ; 1954, *Phys. Rev.*, **95**, 640.
KURATH, D., 1952, *Phys. Rev.*, **88**, 804 ; 1957, *Ibid.*, **106**, 975.
KUZNETSOV, V. M. 1957, *Zh. eksper. teor. Fiz.*, **32**, 1001.
LANE, A. M., and ELLIOTT, J. P., 1958, *Handbuch der Physik*, Ed. S. Flugge (Berlin : Springer).
LANE, A. M., and THOMAS, R. G., 1958, *Rev. Mod. Phys.*, **30**, 257.
LEMONICK, A., CORNWELL, R. G., and ALMQVIST, E., 1959, *Bull. Amer. Phys. Soc.*, Ser. II, **4**, 219.
LEWIS, I. A. D., and WELLS, F. H., 1954, *Millimicrosecond Pulse Techniques* (London : Pergamon Press).
LILLIE, A. B., BONNER, T. W., and CONNER, J. R., 1952, *Phys. Rev.*, **88**, 473.
LINDHARD, J., and SCHARFF, M., 1953, *K. Danske Vidensk. Selsk., Mat-Fys. Medd.*, **27**, No. 15.
LITHERLAND, A. E., 1959, *Bull. Amer. Phys. Soc.*, Ser. II, **4**, 58.
LITHERLAND, A. E., ALMQVIST, E., BATCHELOR, R., and GOVE, H. E., 1959, *Phys. Rev. Lett.*, **2**, 104.
LITHERLAND, A. E., ALMQVIST, E., BROMLEY, D. A., GOVE, H. E., and FERGUSON, A. J., 1957, *Bull. Amer. Phys. Soc.*, Ser. II, **2**, 51.
LITHERLAND, A. E., and BROMLEY, D. A., 1960, *Nuclear Instruments*, in the press.
LITHERLAND, A. E., and FERGUSON, A. J., 1960, *Canad. J. Phys.*, in the press.

LITHERLAND, A. E., and GOVE, H. E., 1958, *Bull. Amer. Phys. Soc.*, Ser. II, **3**, 200.
LITHERLAND, A. E., and MCCALLUM, G., 1960, *Canad. J. Phys.*, in the press.
LUBIRSKY, P. I., MESCHERYAKOV, M. G., and KHRENINA, T. I., 1947, *C.R. Acad. Sci. U.R.S.S.*, **55** (2), 117.
MADANSKY, L., and OWEN, G. E., 1955, *Phys. Rev.*, **99**, 1608.
MANN, L. G., MEYERHOF, W. E., and WEST, H. J., 1953, *Phys. Rev.*, **92**, 1481.
MARION, J. B., 1960, *Phys. Rev.*, in the press.
MARION, J. B., and BONNER, T. W., 1960, *Fast Neutron Physics*, Chap. 5, Ed. J. B. Marion and J. L. Fowler (New York : Academic Press).
MASSEY, B. J., 1957, *Oak Ridge National Laboratory Report* ONRL 2325.
MASSEY, H. S. W., 1950, *Negative Ions* (Cambridge : University Press).
MATTAUCH, J., WALDMANN, L., BIERI, R., and EVERLING, F., 1956 a, *Z. Naturf.*, **11a**, 525 ; 1956 b, *Annual Reviews of Nuclear Science*, Vol. 6 (Palo Alto, California ; Annual Reviews).
MAYER, J., and GOSSICK, B., 1956, *Rev. Sci. Instrum.*, **27**, 407.
MCCARTHY, I. E., 1959, *Nuclear Physics*, **11**, 574.
MCKAY, K. G., 1951, *Phys. Rev.*, **84**, 829.
MCKENZIE, J. M., and BROMLEY, D. A., 1959, *Phys. Rev. Lett.*, **2**, 303 ; 1959 a, *Proceedings of the Conference on Transistors and Semi-conducting Devices* (*Supplement to Proc. I.E.E.* (*London*)), in the press.
METZGER, F. R., 1955, *J. Franklin Inst.*, **260**, 239 ; 1956, *Phys. Rev.*, **103**, 983.
METZGER, F. R., SWANN, C. P., and RASMUSSEN, V. K., 1958, *Phys. Rev.*, **110**, 906.
METZGER, F. R., 1958, *Proc. Rehovoth Conf. Nucl. Structure*, p. 560, Ed. H. J. Lipkin (Amsterdam : North-Holland).
MIDDLETON, R., and TAI, C. T., 1952, *Proc. Phys. Soc.* A, **65**, 752.
MILEIKOWSKY, C., 1954, *Ark. Fys.*, **7**, 33.
MILLER, P. D., and PHILLIPS, G. C., 1958, *Phys. Rev.*, **112**, 2048.
MILLER, W. F., REYNOLDS, J., and SNOW, W. J., 1958, *Argonne National Laboratory Report* ANL–5902.
MOAK, C. D., 1953, *Phys. Rev.*, **92**, 383.
MOAK, C. D., GALONSKY, A., TRAUGHBER, R. L., and JONES, C. M., 1958, *Phys. Rev.*, **110**, 1369.
MOAK, C. D., GOOD, W. M., and KUNZ, W. E., 1954 a, *Phys. Rev.*, **95**, 641 ; 1954 b, *Ibid.*, **96**, 1363.
MOAK, C. D., REESE, H., and GOOD, W. M., 1951, *Nucleonics*, **9**, 318.
MOAK, C. D., and WISSEMAN, W. R., 1956, *Phys. Rev.*, **101**, 1326.
MORPURGO, G., 1958 a, *Phys. Rev. Lett.*, **1**, 68 ; 1958 b, *Phys. Rev.*, **110**, 721.
MURRAY, R. B., 1958, *Nuclear Instruments*, **2**, 237.
NAGGIAR, M. V., ROCLAWSKI-CONJEAUD, M., SZTEINSZNAIDER, D., and THIRION, M. J., 1956, *J. Phys. Radium*, **17**, 561.
NEILER, J. H., 1956, *Bull. Amer. Phys. Soc.*, Ser. II, **1**, 70 ; see also W. M. Good, J. H. Neiler and J. H. Gibbon, 1958, *Phys. Rev.*, **109**, 926.
NEUDACHIN, V. G., 1959, *Soviet Phys., JETP*, **35** (8), 815.
NEWNS, H. C., 1952, *Proc. Phys. Soc.* A, **65**, 916.
NEWTON, T. D., 1950, *Phys. Rev.*, **78**, 490.
NORTON, F. J., 1957, *J. Appl. Phys.*, **28**, 34.
Nuclear Data Sheets, 1959, ed. by C. L. McGinnis *et al.* (Washington : National Bureau of Standards).
OWEN, G. E., and MADANSKY, L., 1957, *Phys. Rev.*, **99**, 1608 ; 1957 a, *Ibid.*, **105**, 1766 ; 1958, *Amer. J. Phys.*, **26**, 260.
OWEN, G. E., MADANSKY, L., and EDWARDS, S., 1959, *Phys. Rev.*, **113**, 1575.
OWEN, R. B., 1958, *Harwell Report* AERE EL/R 2712, *I.R.E. Transactions on Nuclear Science* NS–5, No. 3, 198.
PESHKOV, V. P., 1956, *Zh. eksper. teor. Fiz.*, **30**, 850.
PETIAU, G., 1955, *La theorie des fonctions de Bessel*, p. 70 (Paris : Centre National de la Recherche Scientifique).

PHILLIPS, J. A., 1955, *Phys. Rev.*, **97**, 404.
PHILLIPS, G. C., and MILLER, P., 1957, *Bull. Amer. Phys. Soc.*, Ser. II, **2**, 28.
POOL, M. L., 1952, *Physica*, **18**, 304.
POOL, M. L., and KUNDU, D. N., 1953, *Phys. Rev.*, **91**, 462.
PORTER, W. S., ROTH, B., and JOHNSON, J. J., 1958, *Phys. Rev.*, **111**, 1578.
POVH, B., 1958, *Bull. Amer. Phys. Soc.*, Ser. II, **3**, 406 ; 1959, *Phys. Rev.*, **114**, 1114.
PRICE, W. J., 1958, *Nuclear Radiation Detection* (New York : McGraw-Hill).
RAO, D. A. A. S. N., YOUNG, T. E., SPENCER, R. R., BECKNER, E. H., and PHILLIPS, G. C, 1959, *Bull. Amer. Phys. Soc.*, Ser. II, **4**, 96.
REDLICH, M. G., 1954, *Phys. Rev.*, **95**, 448 ; 1958, *Ibid.*, **110**, 468.
ROLLIN, B. V., and HATTON, J., 1948, *Phys. Rev.*, **74**, 506.
ROSE, M. E., 1953, *Phys. Rev.*, **91**, 610 ; 1958, *Quantum Theory of Angular Momentum* (New York : John Wiley).
ROSSI, B. B., and STAUB, H. H., 1949, *Ionization Chambers and Counters* (New York : McGraw-Hill).
ROTBLAT, J., 1950, *Progress in Nuclear Physics*, Vol. 1, Ed. O. R. Frisch (New York : Academic Press).
SAYRES, A., and WU, C. S., 1958, *Rev. Sci. Instrum.*, **28**, 758.
SCHECTER, L., CRANDALL, W. E., MILLBURN, G. P., and ISE, J., 1955, *Phys. Rev.*, **97**, 184.
SCHIFFER, J. P., BONNER, T. W., DAVIS, R. H., and PROSSER, F. W, 1956 a, *Bull. Amer. Phys. Soc.*, Ser. II, **1**, 324 ; 1956 b, *Phys. Rev.*, **104**, 1064.
SCHUTTE, O. F., ZUCKER, A., and WATSON, W. W., 1950, *Rev. Sci. Instrum.*, **21**, 1016.
SEVERIENS, J. C., and HANNA, S. S., 1956, *Phys. Rev.*, **104**, 1612.
SHARP, W. T., GOVE, H. E., and PAUL, E. B., 1953, *Chalk River Report* TPI–70.
SHARP, W. T., KENNEDY, J. M., SEARS, B. J., and HOYLE, M. G., 1954, *Chalk River Report* CRT–556.
SOLLER, T., FAIRBANK, W. M., and CROWELL, A. D., 1953, *Phys. Rev.*, **91**, 1058.
SPENCER, R. R., and PHILLIPS, G. C., 1959, *Bull. Amer. Phys. Soc.*, Ser. II, **4**, 95.
STAUB, H. H., 1953, *Experimental Nuclear Physics*, Vol. 2, Ed. E. Segrè (New York : Wiley).
STOKES, R. H., NORTHROP, J. A., and BOYER, K., 1958, *Rev. Sci. Instrum.*, **29**, 61.
SWEETMAN, D. R., 1955, *Phil. Mag.*, **46**, 358 ; 1958, *Bull. Amer. Phys. Soc.*, Ser. II, **3**, 186.
TEICHMANN, T., and WIGNER, E. P., 1952, *Phys. Rev.*, **87**, 123.
TEMMER, G. M., and HEYDENBURG, N., 1958, *Phys. Rev.*, **111**, 1303.
THONEMANN, P., MOFFATT, J., ROAF, D., and SANDERS, J. H., 1948, *Proc. Phys. Soc.*, **61**, 483.
TITTERTON, E. W., 1958, *Progress in Nuclear Physics*, Vol. 4, Ed. O. R. Frisch (London : Pergamon Press).
TUBIS, A., 1958, *Los Alamos Laboratory Report*, LA 2150.
TUCK, J. L., ARNOLD, W. R., PHILLIPS, J. A., SAWYER, G. A., and STOVALL, E. J., 1952, *Phys. Rev.*, **88**, 159A.
TURNER, C. M., 1959, *University of California Radiation Laboratory Report* UCRL–352.
TURNER, C. M., and BLOOM, S. D., 1958, *Rev. Sci. Instrum.*, **29**, 480.
VAN RENNES, A. B., 1952, *Nucleonics*, **10**, 20 (July); *Ibid.*, **10**, 23 (Aug.); *Ibid.*, **10**, 32 (Sept.); *Ibid.*, **10**, 51 (Oct.).
VISSCHER, W. M., and FERRELL, R. A., 1955, *Univ. of Maryland Technical Report* No. 19, September ; 1957, *Phys. Rev.*, **107**, 781.
VOGT, E., 1959, *Nuclear Reactions*, Chap. V, Ed. P. M. Endt and M. Demeur (Amsterdam : North-Holland).
WAPSTRA, A. H., 1955, *Physica*, **21**, 367 ; 1955 a, *Ibid.*, **21**, 385.
WARBURTON, E. K., and ROSE, H. J., 1958, *Phys. Rev.*, **109**, 1109.
WAY, K., 1958, *Nuclear Data Cards* (Washington : N.R.C.).
WEGNER, H. E., and HALL, W. S., 1958 a, *Bull. Amer. Phys. Soc.*, Ser. II, **3**, 338 ; 1958 b, *Rev. Sci. Instrum.*, **29**, 1100 ; 1959, *Bull. Amer. Phys. Soc.*, Ser. II, **4**, 95.
WEINMAN, J. A., and CAMERON, J. R., 1956, *Rev. Sci. Instrum.*, **27**, 288.
WERNER, A., 1956, *Nuclear Physics*, **1**, 9.
WEST, D., 1953, *Progress in Nuclear Physics*, Vol. 3, Ed. O. Frisch (London : Pergamon Press).
WHALING, W., 1958, *Handbuch der Physik*, **34**, 193, Ed. S. Flugge (Berlin : Springer).

WHITTIER, A. C., 1954, *Canad. J. Phys.*, **32**, 275.
WILKINSON, D. H., 1950, *Ionization Chambers and Counters* (Cambridge: University Press); 1958, *Proc. Rehovoth Conf. Nucl. Structure* 1957, Ed. H. J. Lipkin (Amsterdam: North-Holland).
WILKINSON, D. H., TOPPEL, B. J., and ALBURGER, D., 1956, *Phys. Rev.*, **101**, 673.
WILSON, J. G., 1951, *The Principle of Cloud Chamber Technique* (Cambridge: University Press).
WILZBACH, K. E., VAN DYKEN, A. R., and KAPLAN, L., 1953, *Argonne National Laboratory Report* ANL 5143.
WINN, M. M., 1954, *Proc. Phys. Soc.* A, **67**, 949.
WOLICKI, E. A., GEER, E., HOLMGREN, H. D., and JOHNSTON, R. L., 1956, *Bull. Amer. Phys. Soc.*, Ser. II, **1**, 196.
WOLICKI, E. A., HOLMGREN, H. D., JOHNSTON, R. L., and GEER, E., 1956 a, *Bull. Amer. Phys. Soc.*, Ser. II, **1**, 325.
WOLICKI, E. A., JASTROW, R., and BROOKS, H. B., 1956 b, *Naval Research Lab. Report* NRL-4833.
WOLICKI, E. A., JOHNSTON, R. L., HOLMGREN, H. D., HORNYAK, W. F., and LEVESQUE, R., 1959, *Bull. Amer. Phys. Soc.*, Ser. II, **4**, 232.
WOLICKI, E. A., HOLMGREW, H. D., JOHNSTON, R. L., and GEER-ILLSLEY, E., 1959 a, *United States Naval Research Laboratory Progress Report*, August issue.
WRIGHT, G. T., 1956, *Proc. Phys. Soc.* B, **49**, 358.
YARNELL, J. L., LOVBERG, R. H., and STRATTON, W. R., 1952, *Phys. Rev.*, **88**, 158A; 1953, *Ibid.*, **90**, 292.

APPENDIX

In addition to the topics presented in this review, Chalk River Report CRP–881 (Bromley and Almqvist 1959) includes a survey of relevant experimental and analytic techniques. The following is an excerpt from the index of this report covering this material.

Experimental Techniques

A.1. Historical survey

A.2. Production of ^{3}He

A.3. Handling of ^{3}He in accelerators
- A.3.1. General discussion
- A.3.2. Gas and ion beam contaminants

A.4. Gas recovery systems
- A.4.1. Chalk River system
- A.4.2. Gas purification

A.5. Gas recirculating systems
- A.5.1. Introduction
- A.5.2. Mercury diffusion pump systems
- A.5.3. Mechanical pump systems

A.6. Ion sources for ^{3}He
- A.6.1. Introduction
- A.6.2. Chalk River ion source system
- A.6.3. Sources of negative or neutral He ions
- A.6.4. Ion beam pulsing

A.7. Targets
A.7.1. Introduction
A.7.2. Target contaminants

A.8. Detectors
A.8.1. Introduction
A.8.2. Charged particle detection
A.8.2.1. Gas ionization detectors
A.8.2.2. Junction detectors
A.8.2.3. Scintillation detectors
A.8.2.4. Magnetic spectroscopy
A.8.2.5. Photographic plate detectors
A.8.3. Neutron detection
A.8.3.1. Reaction counters
A.8.3.2. Scintillation detectors
A.8.3.3. Photographic plate detectors
A.8.3.4. Recoil counters
A.8.3.5. Activation detectors
A.8.4. Gamma radiation detection
A.8.4.1. Scintillation spectrometers
A.8.4.2. Proportional counters
A.8.4.3. Pair spectrometers
A.8.4.4. Crystal diffraction spectrometers

A.9. Detector systems
A.9.1. Fast–slow coincidence systems
A.9.2. Angular correlation techniques
A.9.3. Two-dimensional pulse height analysis
A.9.4. Particle identification systems
A.9.5. Compton spectrometer–polarimeter
A.9.6. Hoogenboom spectrometer
A.9.7. Three crystal pair spectrometer
A.9.8. Lifetime measurements, $\tau > 10^{-3}$ second
A.9.9. Lifetime measurements, $\tau < 10^{-3}$ second
A.9.10. Neutron counter ratio systems
A.9.11. Neutron–gamma differentiation systems
A.9.12. Time-of-flight systems
A.9.13. Resonance scattering systems